



ENCYCLOPEDIA OF

Physical Science AND Technology

THIRD EDITION

Chemical Engineering



Table of Contents
(Subject Area: Chemical Engineering)

Article	<i>Authors</i>	Pages in the Encyclopedia
Absorption (Chemical Engineering)	<i>James R. Fair and Henry Z. Kister</i>	Pages 1-25
Adsorption (Chemical Engineering)	<i>Douglas M. Ruthven</i>	Pages 251-271
Aerosols	<i>G. M. Hidy</i>	Pages 273-299
Batch Processing	<i>Narses Barona</i>	Pages 41-56
Catalysis, Industrial	<i>Bruce E. Leach</i>	Pages 491-500
Catalyst Characterization	<i>Robert J. Farrauto and Melvin C. Hobson</i>	Pages 501-526
Chemical Process Design, Simulation, Optimization,	<i>B. Wayne Bequette and Louis P. Russo</i>	Pages 751-766
Coherent Control of Chemical Reactions	<i>Robert J. Gordon and Yuichi Fujimura</i>	Pages 207-231
Cryogenic Process Engineering	<i>Klaus D. Timmerhaus</i>	Pages 13-36
Crystallization Processes	<i>Ronald W. Rousseau</i>	Pages 91-119
Distillation	<i>M. R. Resettarits and M. J. Lockett</i>	Pages 547-559
Electrochemical Engineering	<i>Geoffrey Prentice</i>	Pages 143-159
Fluid Dynamics (Chemical Engineering)	<i>Richard W. Hanks</i>	Pages 45-70
Fluid Mixing	<i>J. Y. Oldshue</i>	Pages 79-104
Heat Exchangers	<i>Kenneth J. Bell</i>	Pages 251-264
High-Pressure Synthesis (Chemistry)	<i>R. H. Wentorf, Jr. and R. C. DeVries</i>	Pages 365-379

Mass Transfer and Diffusion	<i>E. L. Cussler</i>	Pages 171-180
Membranes, Synthetic, Applications	<i>Eric K. Lee and W. J. Koros</i>	Pages 279-344
Metalorganic Chemical Vapor Deposition	<i>Russell D. Dupuis</i>	Pages 495-511
Pollution Prevention from Chemical Processes	<i>Kenneth L. Mulholland and Michael R. Overcash</i>	Pages 593-609
Pulp and Paper	<i>Raymond A. Young, Robert Kundrot and David A. Tillman</i>	Pages 249-265
Reactors in Process Engineering	<i>Gary L. Foutch and Arland H. Johannes</i>	Pages 23-43
Solvent Extraction	<i>Teh C. Lo and M. H. I. Baird</i>	Pages 341-362
Surfactants, Industrial Applications	<i>Tharwat F. Tadros</i>	Pages 423-438
Synthetic Fuels	<i>Ronald F. Probststein and R. Edwin Hicks</i>	Pages 467-480
Thermal Cracking	<i>B. L. Crynes, Lyle F. Albright and Loo-Fung Tan</i>	Pages 613-626



Absorption (Chemical Engineering)

James R. Fair

University of Texas at Austin

Henry Z. Kister

Fluor-Daniel Corp.

- I. Absorption in Practice
- II. Principles of Absorption
- III. Models for Absorption Equipment
- IV. Absorber Design

GLOSSARY

Absorption factor Ratio of liquid to gas flow rate divided by the slope of the equilibrium curve.

Films Regions on the liquid and gas sides of the interface in which fluid motion is considered slow and through which material is transported by molecular diffusion alone.

Gas solubility Quantity of gas dissolved in a given quantity of solvent at equilibrium conditions.

Hatta number Ratio of the maximum conversion of reacting components into products in the liquid film to the maximum diffusion transport through the liquid film.

Height of a transfer unit Vertical height of a contactor required to give a concentration change equivalent to one transfer unit.

Ideal stage Hypothetical device in which gas and liquid are perfectly mixed, are contacted for a sufficiently long

period of time so that equilibrium is obtained, and are then separated.

Inerts Gas components that are not absorbed by the liquid.

Interface Surface separating the liquid from the gas. Equilibrium is assumed to exist at this surface.

LPG Liquified petroleum gas.

Lean gas Gas leaving the absorber, containing the inerts and little or no solute.

Lean solvent Solvent entering the absorber, containing little or no solute.

Mass transfer coefficient Quantity describing the rate of mass transfer per unit interfacial area per unit concentration difference across the interface.

Number of transfer units Parameter that relates the change in concentration to the average driving force. It is a measure of the ease of separation by absorption.

Operating line Line on the y - x diagram that represents the locus of all the points obeying the component material balance.

Rich gas Gas entering the absorber, containing both the inerts and solutes.

Rich solvent Solvent leaving the absorber, which contains solute removed from the feed gas.

Slope of equilibrium curve Ratio of the change of the solute concentration in the gas to a given change in solute concentration in the liquid when the solvent and solute are at equilibrium and when solute concentrations are expressed as mole fractions.

Solute(s) Component(s) absorbed from the gas by the liquid

Solvent Dissolving liquid used in an absorption process.

Stripping (or desorption) Process in which the absorbed gas is removed from the solution.

y - x diagram Plot in which the solute mole fraction in the gas is plotted against the solute mole fraction in the liquid.

ABSORPTION is a unit operation in which a gas mixture is contacted with a suitable liquid for the purpose of preferentially dissolving one or more of the constituents of the gas. These constituents are thus removed or partially removed from the gas into the liquid. The dissolved constituents may either form a physical solution with the liquid or react chemically with the liquid. The dissolved constituents are termed *solutes*, while the dissolving liquid is termed the *solvent*. When the concentration of solute in the feed gas is low, the process is often called *scrubbing*.

The inverse operation, called stripping, desorption, or regeneration, is employed when it is desirable to remove the solutes from the solvent in order to recover the solutes or the solvent or both.

I. ABSORPTION IN PRACTICE

A. Commercial Application

Absorption is practiced for the following purposes:

1. Gas purification, for example, removal of pollutants from a gas stream.
2. Production of solutions, for example, absorption of hydrogen chloride gas in water to form hydrochloric acid.
3. Product recovery, for example, absorption of liquified petroleum gases (LPG) and gas olines from natural gas.
4. Drying, for example, absorption of water vapor from a natural gas mixture.

Some common commercial applications of absorption are listed in [Table I](#).

B. Choice of Solvent for Absorption

If the main purpose of absorption is to generate a specific solution, as in the manufacture of hydrochloric acid, the solvent is specified by the nature of the product. For all other purposes, there is some choice in selecting the absorption liquid. The main solvent selection criteria are as follows:

1. Gas solubility. Generally, the greater the solubility of the solute in the solvent, the easier it is to absorb the gas, reducing the quantity of solvent and the equipment size needed for the separation. Often, a solvent that is chemically similar to the solute or that reacts chemically with the solute will provide high gas solubility.
2. Solvent selectivity. A high selectivity of the solvent to the desired solutes compared with its selectivity to other components of the gas mixture lowers the quantity of undesirable components dissolved. Application of a solvent of higher selectivity reduces the cost of downstream processing, which is often required to separate out the undesirable components.
3. Volatility. The gas leaving the absorber is saturated with the solvent. The more volatile the solvent is, the greater are the solvent losses; alternatively, the more expensive are the down-stream solvent separation facilities required to reduce the losses.
4. Effects on product and environment. For example, toxic solvents are unsuitable for food processing; noxious solvents are unsuitable when the gas leaving the absorber is vented to the atmosphere.
5. Chemical stability. Unstable solvents may be difficult to regenerate or may lead to excessive losses due to decomposition.
6. Cost and availability. The less expensive is the solvent, the lower is the cost of solvent losses. Water is the least expensive and most plentiful solvent.
7. Others. Noncorrosiveness, low viscosity, nonflammability, and low freezing point are often desirable properties.

C. Absorption Processes

Absorption is usually carried out in a countercurrent tower, through which liquid descends and gas ascends. The tower may be fitted with trays, filled with packing, or fitted with sprays or other internals. These internals provide the surface area required for gas-liquid contact.

A schematic flow diagram of the absorption-stripping process is shown in [Fig. 1](#). Lean solvent enters at the top

TABLE I Common Commercial Applications of Gas Absorption

Type of plant	Feed gas	Solutes	Solvent	Commercial purpose	Stripping practice
Refineries, natural gas plants, petrochemical plants, coal processing plants, hydrogen plants	Refinery gas, natural gas, LPG towns gas, coal gas, hydrogen reformer gas	Hydrogen sulfide, carbon dioxide, mercaptans	Ethanolamines, alkaline solutions, potassium carbonate	Gas purification for downstream processing or to achieve product specifications	Stripping practiced when using ethanolamines or carbonate for the purpose of solvent recovery and recycle; stripping normally not practiced when using an alkaline solution
Combustion plants	Combustion gases	Sulfur dioxide	Water, alkaline solutions	Pollutant removal	Stripping normally not practiced
Natural gas plants	Natural gas	Water	Glycol	Gas drying for further processing or to achieve product specification	Stripping practiced for solvent recovery
Refineries, natural gas plants, petrochemical plants	Gas stream containing mostly hydrogen, methane, and light gases as well as some LPG and gasolines	LPG, gasolines	Kerosene, diesel, gas oil, other refinery oils	Product recovery of LPG, gasolines	Stripping practiced for LPG and gasoline recovery
Coke ovens	Coke oven gas	Benzene, toluene	Heavy oil	By-product recovery	Stripping practiced to recover the by-product
Sulfuric acid	Sulfur trioxide mixed with oxygen and nitrogen	Sulfur trioxide	Sulfuric acid, oleum	Sulfuric acid manufacture	Stripping not practiced
Nitric acid	Nitrogen dioxide mixed with nitrogen oxide, oxygen, nitrogen	Nitrous oxides	Nitric acid, water	Nitric acid manufacture	Stripping not practiced
Carbon dioxide	Combustion gases, kiln gases	Carbon dioxide	Carbonate, bicarbonate solution	Carbon dioxide production	Stripping practiced to recover carbon dioxide
Hydrochloric acid	By-products of chlorination reaction	Hydrogen chloride	Hydrochloric acid, water	Hydrochloric acid production	Stripping not practiced
Soda ash (sodium carbonate), mineral processing	Combustion gases, lime-kiln gases	Carbon dioxide	Ammonia solution	Ammonium bicarbonate production, ammonium carbonate production	Stripping not practiced
Soda ash (sodium carbonate), mineral processing	Waste gases, ammonia makeup	Ammonia	Brine solution	Production of ammonium hydroxide for ammonium bicarbonate production	Stripping not practiced
Hydrogen cyanide	Tail gases, ammonia, hydrogen cyanide	Ammonia	Sulfuric acid	Ammonia removal while producing ammonium sulfate by-product	Stripping not practiced
Hydrogen cyanide, acrylonitrile	Hydrogen cyanide, tail gases, acrylonitrile	Hydrogen cyanide acrylonitrile	Water	Separation of hydrogen cyanide and acrylonitrile from tail gases	Stripping is practiced to recover hydrogen cyanide and acrylonitrile from water

Continues

TABLE I (continued)

Type of plant	Feed gas	Solutes	Solvent	Commercial purpose	Stripping practice
Ethylene oxide, glycol	Reactor effluent	Ethylene oxide	Water	Ethylene oxide recovery	Stripping is practiced to recover ethylene oxide from the solution
Ketones from alcohol	Hydrogen, ketones	Ketones	Water	Ketone–hydrogen separation	Stripping is practiced to recover ketones from the solution
Maleic anhydride	Reactor effluent	Maleic anhydride separation	Water	Maleic anhydride from reactor gases	Stripping is practiced to remove water from the maleic acid formed in the absorption process, converting it back to maleic anhydride
Isoprene	Reactor effluent	Isoprene, C ₄ 's, C ₅ 's	Heavy oil	Separation of C ₄ 's, C ₅ 's, and isoprene from light gases	Stripping is practiced to recover the solute and regenerate the oil for recycling to the absorber
Urea	Reactor effluent	CO ₂ , NH ₃	Water	Formation of ammonium carbonate solution, which is recycled to the reactor	Stripping not practiced

of the absorber and flows downward through the internals. Rich gas enters at the bottom of the absorber and flows upward through the internals. The liquid and gas are contacted at the absorber internals, and the solute is absorbed by the solvent. Overhead product from the absorber is the

solute-free lean gas, and bottom product is the rich solvent, which contains the absorbed solute. The rich solvent then flows to the stripper where the solute is stripped from the rich solvent, this operation being at a higher temperature and/or lower pressure than maintained in the absorber. The solute leaves the stripper as the overhead product, and the solute-free lean solvent leaves the stripper bottom and is recycled to the absorber.

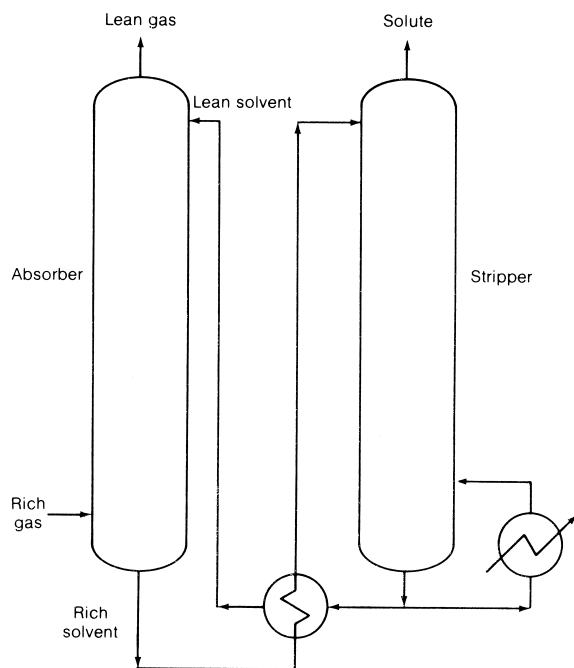


FIGURE 1 Typical schematic absorber–stripper flow diagram.

II. PRINCIPLES OF ABSORPTION

The important fundamental physical principles in absorption are solubility and mass transfer. When a chemical reaction is involved, the principles of reaction equilibria and reaction kinetics are also important.

A. Gas Solubility

At equilibrium, the fugacity of a component in the gas is equal to the fugacity of the same component in the liquid. This thermodynamic criterion defines the relationship between the equilibrium concentration of a component in the gas and its concentration in the liquid. The quantity of gas dissolved in a given quantity of solvent at equilibrium conditions is often referred to as the gas solubility.

Gas solubility data are available from handbooks and various compendia and often show solubility as a function of gas composition, temperature and pressure. A typical

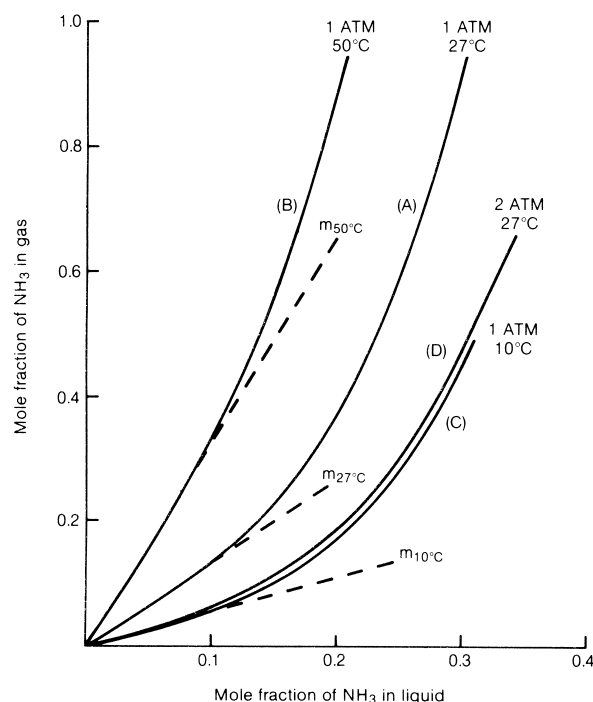


FIGURE 2 Solubility data for NH_3 absorption from air using H_2O . [Data from Perry, R. H., ed. (1985). "Chemical Engineer's Handbook," McGraw-Hill, New York.]

graphical presentation is shown in Fig. 2, where gas composition of a given solute is plotted against liquid composition of the same solute, at equilibrium. Compositions can be represented in various units, such as mole fraction, mole ratio, partial pressure (gas). Figure 2 shows the effect of temperature and pressure on solubility. Solubility is also dependent on whether the solute reacts chemically with the solvent as well as on the nature and amounts of other solutes present.

The equilibrium curve is often approximated linearly,

$$y_A = mx_A \quad (1a)$$

where m is a constant at a given temperature and pressure. This expression is often valid at low concentrations (Fig. 2).

For a solution that is thermodynamically ideal, m is given by "Raoult's law"

$$m = p^{vap}/P \quad (1b)$$

or the ratio of vapor pressure to total pressure. When the gas composition is expressed as partial pressure, the Henry's law coefficient for a given solute is

$$H = p/x \quad (1c)$$

or

$$m = H/P \quad (1d)$$

Henry's law is usually a reasonable approximation at low and moderate concentrations, at constant temperature, and at relatively low pressures (generally less than 5 atm; however, the law may be obeyed at higher pressure at low solubilities).

If a gas mixture containing several components is in equilibrium with a liquid, Henry's law applies separately so long as the liquid is dilute in all the components. If a component is almost insoluble in the liquid, for example, air in water, it has a very high Henry's law constant and a high value of m in Eq. (1). Such a component is absorbed in negligible quantities or by the liquid, and it is often referred to as an *inert component*. The nature and type of the inert component have little effect on the equilibrium curve.

Equilibrium data for absorption are usually available in the literature in three forms:

1. Solubility data, expressed either as mole percent, mass percent, or Henry's law constants
2. Pure-component vapor pressure data
3. Equilibrium distribution coefficients (K values)

To define fully the solubility of a component in a liquid, it is necessary to state the temperature, the partial pressure of the solute in the gas, the concentration of the solute in the liquid, and generally also the pressure.

When gas solubility data are lacking or are unavailable at the desired temperature, they can be estimated using available models. The method of Prausnitz and Shair (1961), which is based on regular solution theory and thus has the limitations of that theory. The applicability of regular solution theory is covered in detail by Hildebrand *et al.* (1970). A more recent model, now widely used, is UNIFAC, which is based on structural contributions of the solute and solvent molecular species. This model is described by Fredenslund *et al.* (1977) and extensive tabulations of equilibrium data, based on UNIFAC, have been published by Hwang *et al.* (1992) for aqueous systems where the solute concentrations are low and the solutions depart markedly from thermodynamic equilibrium.

Perhaps the best source of information on estimating gas solubility is the book by Reid *et al.* (1987), which not only lists the various solubility models but also compares them with a database of experimental measurements.

B. Mass Transfer Principles

The principles of mass transfer determine the rate at which the equilibrium is established, that is, the rate at which the solute is transferred into the solvent.

For a system in equilibrium, no net transfer of material occurs between the phases. When a system is not in equilibrium, diffusion of material between the phases will occur so as to bring the system closer to equilibrium. The departure from equilibrium provides the driving force for diffusion of material between the phases.

The rate of diffusion can be described by the film theory, the penetration theory, or a combination of the two. The most popular description is in terms of a two-film theory. Accordingly, there exists a stable interface separating the gas and liquid. A certain distance from the interface, large fluid motions exist; and these distribute the material rapidly and equally, so that no concentration gradients develop. Next to the interface, however, there are regions in which the fluid motion is slow; in these regions, termed *films*, material is transferred by diffusion alone. At the interface, material is transferred instantaneously, so that the gas and liquid are in equilibrium at the interface. The rate-governing step in absorption is therefore the rate of diffusion in the gas and liquid films adjacent to the interface. The concentration gradient in both phases are illustrated in Fig. 3. Note that y_{Ai} may be higher or lower than x_{Ai} , depending on the equilibrium curve (e.g., Fig. 2); however, y_{Ai} is always lower than y_A , and x_{Ai} is always higher than x_A , or no mass transfer will occur.

1. Dilute Solutions

Applying the diffusion equations to each film and approximating the concentration gradient linearly yields an expression for the mass transfer rates across the films,

$$N_A = k_G(y_A - y_{Ai}) = k_L(x_{Ai} - x_A) \quad (2)$$

This equation states that, for each phase, the rate of mass transfer is proportional to the difference between the bulk concentration and the concentration at the gas–liquid in-

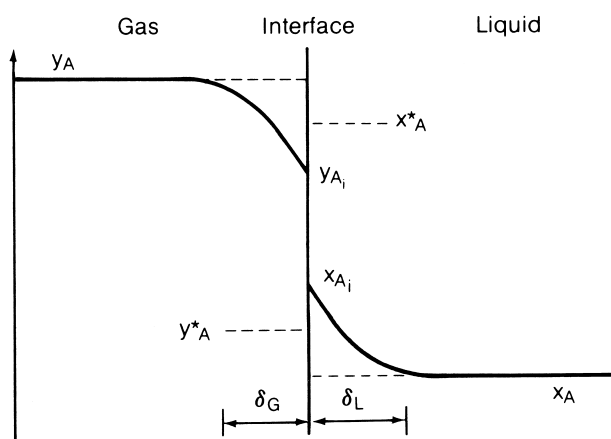


FIGURE 3 Concentration profiles in the vapor and liquid phases near an interface.

terface. Here k_G and k_L are the mass transfer coefficients, and their reciprocals, $1/k_G$ and $1/k_L$ are measures of the resistance to mass transfer in the gas and liquid phases, respectively. Note that the rate of mass transfer in the gas film is equal to that in the liquid film; otherwise, material will accumulate at the interface.

The concentration difference in the gas can be expressed in terms of partial pressures instead of mole fractions, while that in the liquid can be expressed in moles per unit volume. In such cases, an equation similar to Eq. (2) will result. Mole fraction units, however, are generally preferred because they lead to gas mass transfer coefficients that are independent of pressure.

It is convenient to express the mass transfer rate in terms of a hypothetical bulk-gas y_A^* , which is in equilibrium with the bulk concentration of the liquid phase, that is,

$$N_A = K_{OG}(y_A - y_A^*) \quad (3)$$

If the equilibrium curve is linear, as described by Eq. (1), or can be linearly approximated over the relevant concentration range, with an average slope m such that

$$m = (y_A - y_A^*) / (x_A^* - x_A) \quad (4)$$

then Eqs. (2)–(4) can be combined to express K_{OG} in terms of k_G and k_L , as follows:

$$\frac{1}{K_{OG}} = \frac{1}{k_G} + \frac{m}{k_L} \quad (5)$$

Equation (5) states that the overall resistance to mass transfer is equal to the sum of the mass transfer resistances in each of the phases.

The use of overall coefficients is convenient because it eliminates the need to calculate interface concentrations. Note that, theoretically, this approach is valid only when a linear approximation can be used to describe the equilibrium curve over the relevant concentration range. Figure 4 illustrates the application of this concept on an x – y diagram.

For most applications it is not possible to quantify the interfacial area available for mass transfer. For this reason, data are commonly presented in terms of mass transfer coefficients based on a unit volume of the apparatus. Such volumetric coefficients are denoted $k_G a$, $k_L a$ and $K_{OG} a$, where a is the interfacial area per unit volume of the apparatus.

If most of the resistance is known to be concentrated in one of the phases, the resistance in the other phase can often be neglected and Eq. (5) simplified. For instance, when hydrogen chloride is absorbed in water, most of the resistance occurs in the gas phase, and $K_{OG} \approx k_G$. When oxygen is absorbed in water, most of the resistance occurs in the liquid phase, and $K_{OG} \approx k_L/m$.

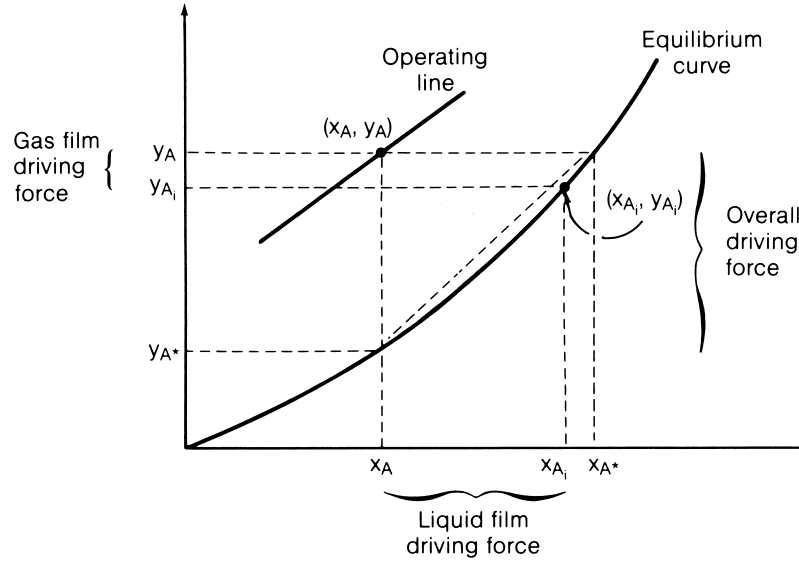


FIGURE 4 Absorption driving forces in terms of the x - y diagram.

2. Concentrated Solutions

Equation (2), derived for dilute solutions, is valid when the flow of solute from the gas to the gas film is balanced by an equal flow of the inert component from the film to the gas; similarly, it requires that the flow of solute from the liquid film to the solvent be balanced by an equal flow of solvent from the liquid into the liquid film. This is a good approximation when both the gas and the liquid are dilute solutions. If either or both are concentrated solutions, the flow of gas out of the film, or the flow of liquid into the film, may contain a significant quantity of solute. These solute flows counteract the diffusion process, thus increasing the effective resistance to diffusion.

The equations used to describe concentrated solutions are derived in texts by [Sherwood *et al.* \(1975\)](#), [Hobler \(1966\)](#), and [Hines and Maddox \(1985\)](#). These reduce to Eqs. (2) and (3) when applied to dilute solutions. These equations are as follows:

$$N_A = k'_G(y_A - y_{Ai})/y_{BM} = k'_L(x_{Ai} - x_A)/x_{BM} = K'_{OG}(y_A - y_A^*)/y_{BM}^*, \quad (6a)$$

where

$$y_{BM} = \frac{(1 - y_A) - (1 - y_{Ai})}{\ln[(1 - y_A)/(1 - y_{Ai})]} \quad (6b)$$

$$x_{BM} = \frac{(1 - x_A) - (1 - x_{Ai})}{\ln[(1 - x_A)/(1 - x_{Ai})]} \quad (6c)$$

$$y_{BM}^* = \frac{(1 - y_A) - (1 - y_A^*)}{\ln[(1 - y_A)/(1 - y_A^*)]} \quad (6d)$$

The terms subscripted BM describe the log-mean solvent or log-mean inert gas concentration difference between the bulk fluid and the interface [Eqs. (6b) and (6c)] or between the bulk fluid and the equilibrium values [Eq. (6d)].

Equation (6a) is analogous to Eqs. (2) and (3). Comparison of these shows that, in concentrated solutions, the concentration-independent coefficients of Eqs. (2) and (3) are replaced by concentration-dependent coefficients in Eq. (6a) such that

$$k_G = k'_G y_{BM} \quad (7a)$$

$$k_L = k'_L x_{BM} \quad (7b)$$

$$K_{OG} = K'_{OG} y_{BM}^* \quad (7c)$$

3. Multicomponent Absorption

The principles involved in multicomponent absorption are similar to those discussed for concentrated solutions. [Wilke \(1950\)](#) developed a set of equations similar to Eq. (6a) to represent this case,

$$N_A = k''_G(y_A - y_{Ai})/y_{fm} = k''_L(x_{Ai} - x_A)/x_{fm} = K''_{OG}(y_A - y_A^*)/y_{fm}^*, \quad (8a)$$

where

$$y_{fm} = \frac{(1 - t_A y_A) - (1 - t_A y_{Ai})}{\ln[(1 - t_A y_A)/(1 - t_A y_{Ai})]} \quad (8b)$$

$$x_{fm} = \frac{(1 - t_A x_A) - (1 - t_A x_{Ai})}{\ln[(1 - t_A x_A)/(1 - t_A x_{Ai})]} \quad (8c)$$

$$y_{fm}^* = \frac{(1 - t_A y_A) - (1 - t_A y_A^*)}{\ln[(1 - t_A y_A)/(1 - t_A y_A^*)]} \quad (8d)$$

$$t_A = \frac{N_A + N_B + N_C + \dots}{N_A} \quad (8e)$$

In a manner similar to the concentrated solutions case, the coefficients in Eqs. (8) can be expressed in terms of the concentration-independent coefficients using relationships similar to those of Eqs. (7), that is,

$$k_G = k_G'' y_{fm} \quad (9a)$$

$$k_L = k_L'' x_{fm} \quad (9b)$$

$$K_{OG} = K_{OG}'' y_{fm}^* \quad (9c)$$

4. Absorption with Chemical Reaction

When the solute is absorbed into a solution containing a reagent that chemically reacts with it, the concentration profile shown in Fig. 3 becomes dependent on the kinetics of the reaction and the concentration of the reacting reagent in the liquid.

Figure 5 shows concentration profiles that commonly occur when solute A undergoes an irreversible second-order reaction with component B, dissolved in the liquid, to give product C,



The rate equation is

$$r_A = k_2 C_A C_B; \quad r_B = b r_A \quad (11)$$

Figure 5 shows that a fast reaction takes place only in the liquid film. In such instances, the dominant mass transfer mechanism is physical absorption and the diffusion model above is applicable, but the resistance to mass transfer in the liquid phase is lower because of the reaction. On the other hand, a slow reaction occurs in the bulk of the liquid, and its rate has little dependence on the resistances to diffusion in either the gas or liquid film. Here the dominant mass transfer mechanism is that of chemical reaction; therefore, this case is considered part of chemical reaction technology, as distinct from absorption technology.

The Hatta number Ha is a dimensionless group used to characterize the speed of reaction in relation to the diffusional resistance to mass transfer,

$$Ha = \frac{\text{max. possible conversion in the liquid film}}{\text{max. diffusional transport through the liquid film}} \\ = \frac{D_A k_2 C_{B0}}{(k_L^\circ)^2} \quad (12)$$

When $Ha \gg 1$, all the reaction occurs in the film, and the process is that of absorption with chemical reaction. As in the case of absorption with no reaction, the main consideration is to provide sufficient surface area for diffusion. On the other hand, when $Ha \ll 1$, all the reaction occurs

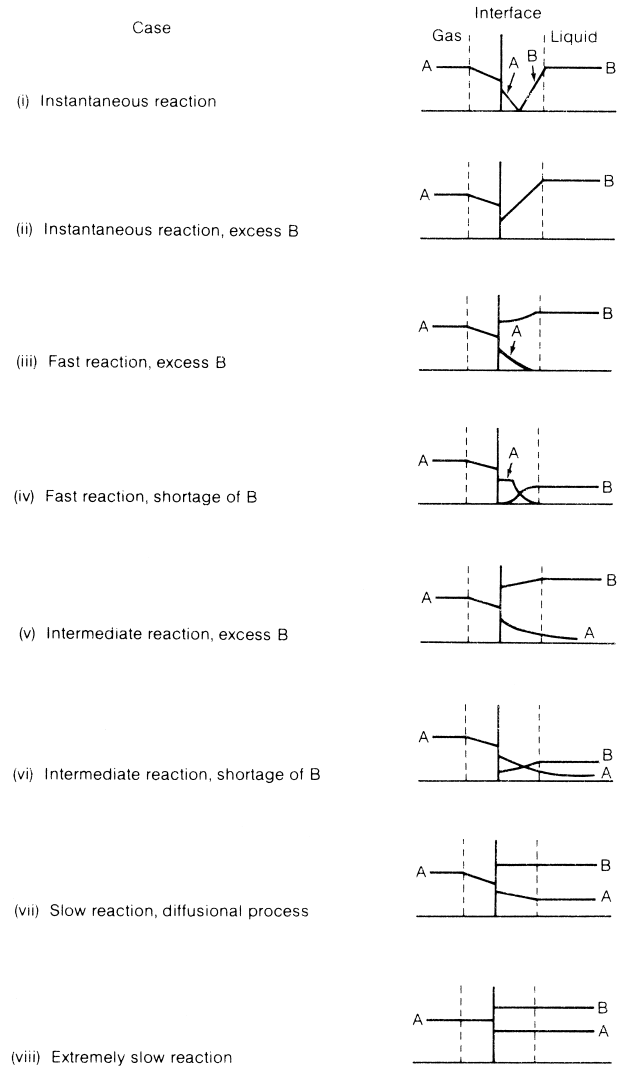


FIGURE 5 Vapor- and liquid-phase concentration profiles near an interface for absorption with chemical reaction.

in the bulk of the liquid, and the contactor behaves as a reactor, not an absorber. Here, the main consideration is providing sufficient liquid holdup for the reaction to take place.

The effect of chemical reaction on rate of absorption is described in terms of an *enhancement factor* ϕ which is used as a multiplier:

$$\phi = k_L / k_L^\circ,$$

where k_L° is the physical mass transfer coefficient.

The enhancement factor can be evaluated from equations originally developed by Van Krevelen and Hoftijzer (1948). A convenient chart based on the equations is shown in Fig. 6. The parameter for the curves is $\phi_x - 1$, where ϕ_x is the enhancement factor as Ha approaches

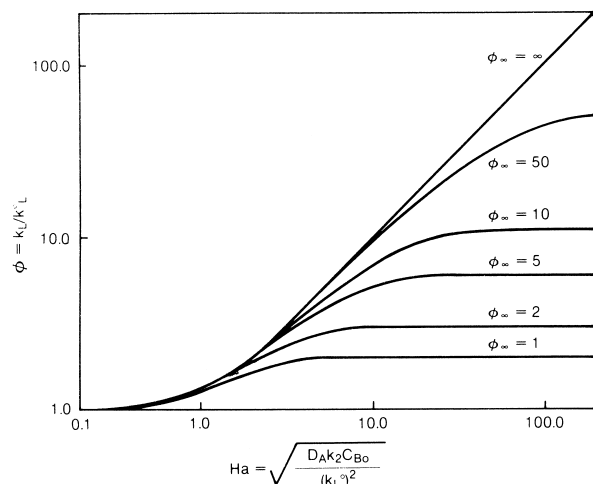


FIGURE 6 Effect of chemical reaction on liquid-phase mass transfer coefficient (assumes bimolecular irreversible reaction). [Data based on Van Krevelen, D. W., and Hofstijzer, P. J. (1948). *Rec. Trav. Chim.* **67**, 563.]

infinity, i.e., when all the solute reacts in the film. Values of ϕ_x were originally based on two-film theory, but a more recent refinement described in Perry's Handbook (Fair, 1997) enables one to make the evaluation in terms of penetration theory, as follows:

$$\phi_\infty = \sqrt{\frac{D_A}{D_B}} + \sqrt{\frac{D_B}{D_A} \left(\frac{C_{B0}}{bC_{Ai}} \right)} \quad (14a)$$

The upper curve of Fig. 6 represents a pseudo-first-order reaction, at which the concentration of B is the same in the film as in the bulk of the liquid. For values of Ha greater than 3, k_L for pseudo-first-order reactions is given by

$$k_L = \sqrt{k_2 C_{B0} D_A} \quad (14b)$$

This discussion applies to an irreversible second-order reaction. For reversible reactions the relationships are more complex and are discussed in the texts by Sherwood *et al.* (1975) and by Danckwerts (1970).

III. MODELS FOR ABSORPTION EQUIPMENT

The principles discussed in Section II describe the equilibrium and mass transfer behavior at a given point. In actual plant equipment, because of the transfer of solute from the gas to the solvent, concentrations change from point to point as the gas and liquid travel through the equipment. These changes cause variations in equilibrium concentrations, driving forces, and mass transfer resistances. The point relationships can be translated into equipment mass

transfer behavior with the aid of material and heat balances. In order to apply these balances, the equipment must be described in terms of a mathematical model.

In this section, the equations are presented for the common types of contactors: differential contactors and stage-wise contactors. The equations are developed for the case of steady-state, countercurrent contacting of liquid and gas with negligible heat effects, with a single-component absorption. Some discussion of extensions to other situations follows.

A. Differential Contactors

1. Material Balances

Differential contactors include packed towers, spray towers, and falling-film absorbers, and are often called *counterflow contactors*. In such devices gas and liquid flow more or less continuously as they move through the equipment.

A material balance over a contactor slice (Fig. 7) gives

$$dG_M = N_A a \, dh \quad (15a)$$

Similarly, a component balance over the same slice gives

$$d(G_M y) = y \, dG_M + G_M \, dy = N_A a \, dh \quad (15b)$$

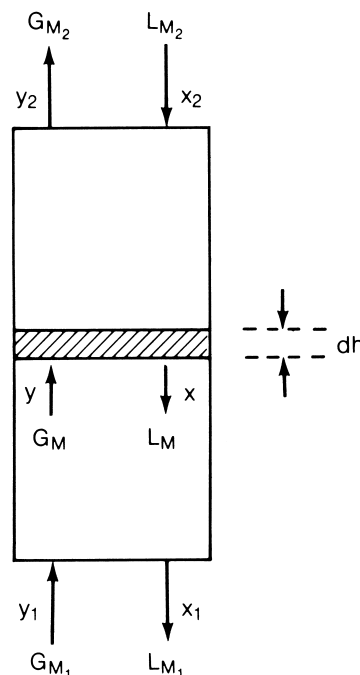


FIGURE 7 Material balance for a differential contactor.

Combining these to eliminate dG_M and integrating gives

$$h_T = \int_{y_2}^{y_1} \frac{G_M dy}{N_A a(1-y)}, \quad (15c)$$

Substituting Eq. (6a) for N_A gives

$$h_T = \int_{y_2}^{y_1} \frac{G_M y_{BM} dy}{k'_G a(1-y)(y-y_1)} \quad (15d)$$

The group $G_M/k'_G a$ is independent of concentration and can be taken out of the integral, giving

$$h_T = \left(\frac{G_M}{k'_G a} \right) \int_{y_2}^{y_1} \frac{y_{BM} dy}{(1-y)(y-y_1)} = H_G N_G \quad (16a)$$

Here N_G is dimensionless and is referred to as the number of gas-phase transfer units; H_G has the dimension of length or height and is referred to as the height of a gas-phase transfer unit.

Here N_G is dimensionless and is called the *number of gas-phase transfer units*; H_G has the dimension of length or height and is referred to as the *height of a gas-phase transfer unit*. As shown in Eq. (16a), the required height of the packed bed h_T is the product of H_G and N_G .

A similar derivation can be carried out in terms of liquid concentrations and flows, giving

$$h_T = H_L N_L = \frac{L_M}{k'_L a} \int_{x_2}^{x_1} \frac{x_{BM} dx}{(1-x)(x_1-x)} \quad (16b)$$

A derivation similar to the preceding one but in terms of the overall mass transfer coefficient K'_{OG} [Eq. (6)] gives

$$h_T = \frac{G_M}{K'_{OG} a} \int_{y_2}^{y_1} \frac{y_{BM}^* dy}{(1-y)(y-y^*)} = H_{OG} N_{OG} \quad (16c)$$

where

$$H_{OG} = G_M / K'_{OG} a \quad (16d)$$

and

$$N_{OG} = \int_{y_2}^{y_1} \frac{y_{BM}^* dy}{(1-y)(y-y^*)} \quad (16e)$$

Equation (16c) is of great practical interest. It is the basis for computing the required packed height for a given separation, and takes into account mass transfer resistances on both sides of the interface. Also, it avoids the need to calculate the interfacial concentrations required for Eqs. (16a) and (16b).

The N_{OG} in Eq. (16e) is termed the *overall number of transfer units*. It is dimensionless and is the ratio of the change of bulk-phase concentration to the average concentration driving force. It is essentially a measure of the ease of separation. The H_{OG} in Eq. (16d) is termed the *overall height of a transfer unit*. It has the dimension of length and defines the vertical height of contactor required to provide a change of concentration equivalent to one transfer unit.

It is therefore a measure of the efficiency of contacting provided by the particular device used in the tower.

Mass transfer data are often expressed in terms of H_G and H_L , and these are used to obtain the value of H_{OG} . The relationship between H_{OG} , H_G , and H_L is obtained by substituting the expressions for H_G , H_L , and H_{OG} in Eqs. (16a)–(16c), together with Eqs. (7a)–(7c), in Eq. (5) to give

$$H_{OG} = \frac{y_{BM}}{y_{BM}^*} H_G + \frac{m G_M}{L_M} \frac{x_{BM}}{y_{BM}^*} H_L \quad (17)$$

2. Dilute Systems

For dilute systems, the x_{BM} , y_{BM} , and $1-y$ terms approach unity, and Eqs. (16e) and (17) can be rewritten

$$N_{OG} = \int_{y_2}^{y_1} \frac{dy}{y-y^*} \quad (18a)$$

$$H_{OG} = H_G + m \frac{G_M}{L_M} H_L \quad (18b)$$

When Henry's law is valid [Eq. (1c)], Eq. (18a) can be analytically integrated; alternatively, the graphical form shown in Fig. 8 can be used for evaluating N_{OG} . Expressions for cases in which the equilibrium curve cannot be linearly approximated are available in several texts, such as Hines and Maddox (1985). Figure 8 shows that the number of transfer units increases with the ratio $m G_M / L_M$. When this ratio increases above unity, the number of transfer units, and therefore column height, rapidly increase;

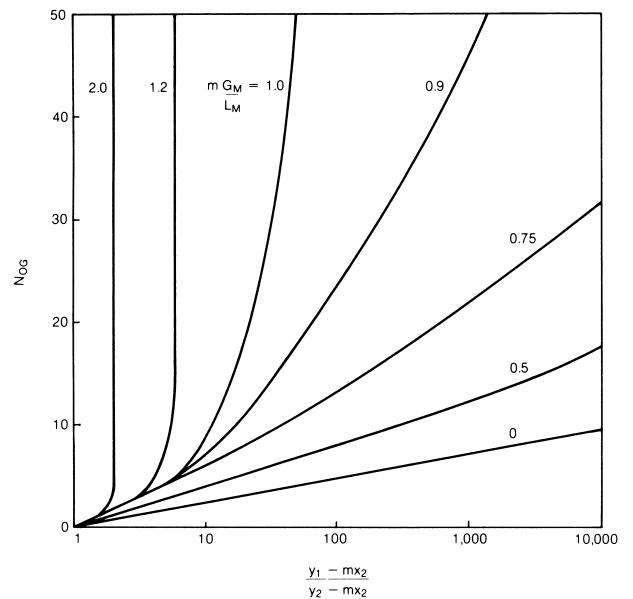


FIGURE 8 Number of overall gas-phase transfer units at constant $m G_M / L_M$.

in such case, a large column height is required to achieve a reasonable level of absorption.

3. Multicomponent Absorption

The above derivations can be extended to multicomponent absorption, making use of Eqs. (8) as described by Hobler and by Sherwood *et al.* (1975) and giving

$$N_{OG} = \int_{y_2}^{y_1} \left(\frac{y_{fm}}{1 - ty} \right) \left(\frac{dy}{y - y^*} \right) \quad (19a)$$

and

$$H_{OG} = \frac{G_M}{K''_{OG} a y_{fm}^*}, \quad (19b)$$

where y_{fm}^* and t are given by Eqs. (8d) and (8e), respectively.

B. Stagewise Contactors

Tray columns and sometimes also packed and spray columns are described in terms of a stage model. An ideal or theoretical stage is hypothetical device in which the gas and liquid are perfectly mixed, contacted for a sufficiently long period of time so that equilibrium is attained, and then separated. The gas leaving the stage is therefore in equilibrium with the liquid leaving the stage. In practice, complete equilibrium can never be attained, since infinite contact time is required to achieve equilibrium. A factor used to account for this nonideality is stage efficiency.

1. Material Balances

An absorber is often modeled as a device that contains a finite number of ideal stages (Fig. 9), with countercurrent flow of vapor and liquid. As the gas rises from stage to stage, it contains less and less of the solute, which is transferred to the solvent.

A material balance can be written for envelope 1 in Fig. 9.

$$L_{M,0}x_0 + G_{M,n}y_n = L_{M,n-1}x_{n-1} + G_{M,1}y_1 \quad (20)$$

The equation can be expressed in terms of the flows entering the absorber, that is, the solute-free solvent entering at the top, and the rich gas, such that

$$y' = yG_M/G'_M \quad (21a)$$

$$x' = xL_M/L'_M \quad (21b)$$

Substituting Eqs. (21a) and (21b) in Eq. (20) gives

$$L'_{M,0}x'_0 + G'_{M,n}y'_n = L'_{M,n-1}x'_{n-1} + G'_{M,1}y'_1 \quad (22)$$

Since the feed flows G'_M and L'_M do not change throughout the contactor,

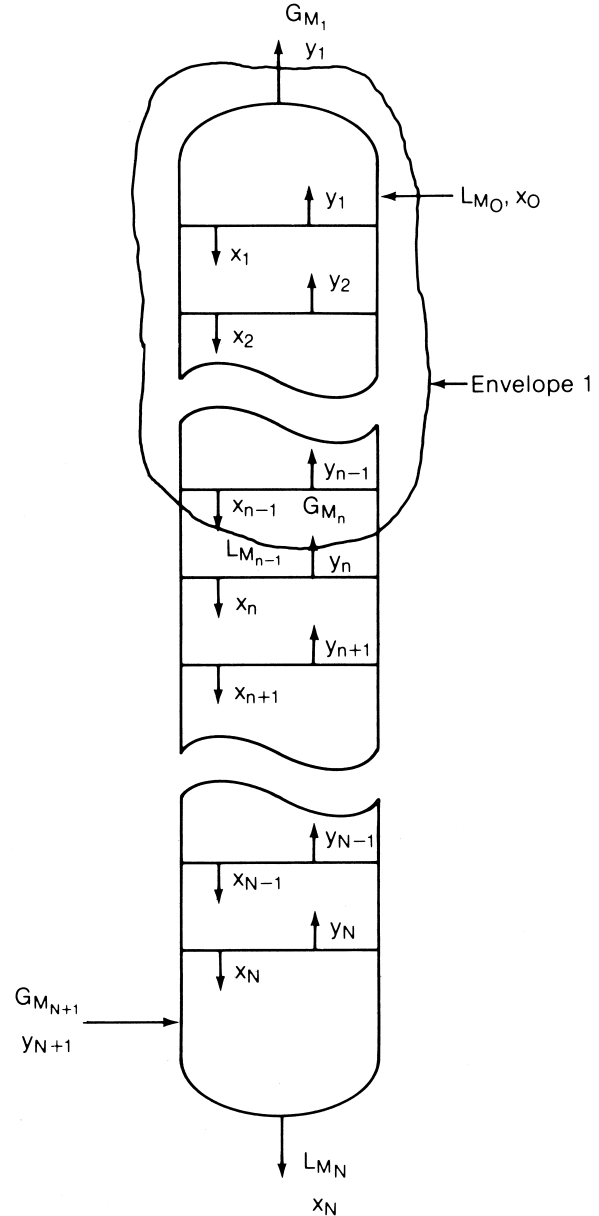


FIGURE 9 Schematic diagram of a stagewise absorber.

$$L'_{M,0} = L'_{M,1} = L'_{M,2} = \cdots = L'_{M,n-1} \\ = L'_{M,n} = \cdots = L'_{M,N} \quad (23a)$$

$$G'_{M,0} = G'_{M,1} = G'_{M,2} = \cdots = G'_{M,n-1} \\ = G'_{M,n} = \cdots = G'_{M,N} \quad (23b)$$

and Eq. (22) can be simplified to give

$$y'_n = \frac{L'_M}{G'_M} x'_{n-1} + \left(y'_1 - \frac{L'_M}{G'_M} x'_0 \right) \quad (24)$$

Equation (24) is an equation of a straight line when plotted on $x - y$ coordinates, with a slope of L'_M/G'_M and an intercept of $y'_1 - L'_M x'_0/G'_M$. This line is often referred to as the *operating line* and is the locus of all the points that obey the stage material balance given by Eq. (20).

For dilute solutions, $L'_M \approx L_M$, $G'_M \approx G_M$, $x' \approx x$, and $y' \approx y$, Eq. (24) simplifies to

$$y_n = \frac{L_M}{G_M} x_{n-1} + \left(y_1 - \frac{L_M}{G_M} x_0 \right) \quad (25)$$

2. Graphical Method

The operating line can be plotted as a straight line on y' versus x' coordinates. The equilibrium curve can also be plotted on the same coordinates (Fig. 10). Each point on the operating line obeys the stage material balance given by Eq. (24); the (x'_n, y'_{n+1}) values of a point on this line give the compositions of the liquid leaving and vapor entering stage n . Each point on the equilibrium curve, given by $y'_n = f(x'_n)$, obeys the equilibrium relationship at stage n ; y'_n is in equilibrium with x'_n .

To obtain the number of ideal stages in the contactor, one starts by plotting the point y'_{N+1} (which is the feed composition of the gas) on the operating line; this defines x'_N . This corresponds to solving the material balance given by Eq. (24) to determine x'_N . Next, one draws a vertical line from the point (x'_N, y'_{N+1}) to the equilibrium curve; this defines y'_N . This corresponds to solving the equilibrium relationship to determine y'_N . From the point

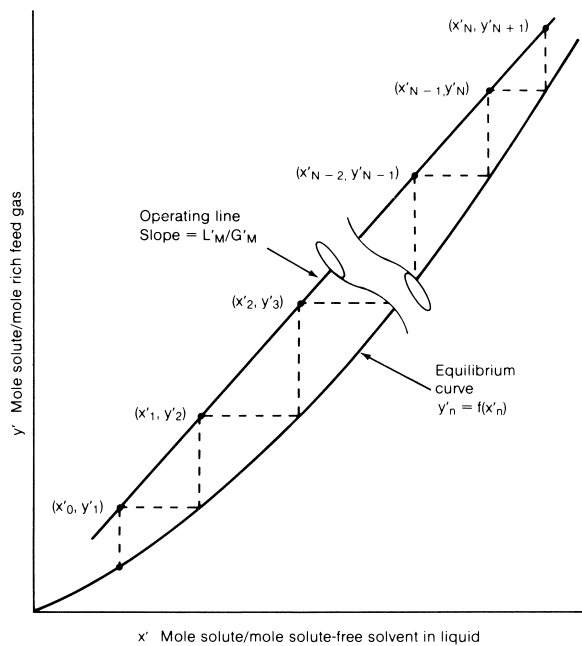


FIGURE 10 Graphic method for stagewise contactors.

(x'_N, y'_N) , one draws a horizontal line back to the operating line, thus solving the material balance to obtain x'_{N-1} . The process is then continued until the top liquid composition x'_0 is reached (Fig. 10). Each step shown on the diagram represents one ideal stage; the number of ideal stages is counted from the diagram.

Often, slightly different coordinates are used for $y-x$ plotting. Instead of plotting y' against x' , one can plot y against x ; in this case, the operating line will be curved. At other times, y' can be expressed in terms of moles of solute per mole of gas leaving the absorber. The construction described above is similar in all these cases.

Rousseau and Staton (1988) extended the $y-x$ diagram plot to situations where a single component A is absorbed into the liquid and instantaneously reacts with a reactive species in the liquid. In this plot, the x axis is the fraction of reactive species in the liquid that has reacted with solute A , while the y axis is the ratio of moles of solute A in the gas to moles of solute-free gas. The equilibrium curve is derived using both Henry's Law constant and the reaction equilibrium constant.

3. Minimum Solvent Rate

When the operating line and the equilibrium curve intersect, an infinite number of stages is required to achieve the separation (Fig. 11). The intersection point is called the *pinch point* and may occur at the bottom (Fig. 11a), at the top (Fig. 11b), or at a tangent point (Fig. 11c). The solvent rate leading to this intersection is the minimum solvent flow required to absorb the specified amount of solute.

Since the top and bottom pinch points shown in Fig. 11 represent intersections of operating and equilibrium lines, they may be predicted analytically. At the top, the lean gas and the lean solvent are in equilibrium, i.e., $y_1 = Kx_0$ (Fig. 9). Similarly, at the bottom the rich gas and the rich solvent are in equilibrium, i.e., $y_{N+1} = mKx_N$. By material balance, the minimum solvent rate can be calculated. Frequently, the pinch occurs at the bottom.

The actual solvent rate specified for the separation must exceed the minimum solvent rate, or an infinite number of stages will be required. For a contactor with a finite number of stages, this means that the separation will not be achieved unless actual solvent rate exceeds the minimum. The higher the solvent rate specified, the greater is the distance between the operating line and the equilibrium curve, and the smaller is the number of stages required.

4. Absorption Factors

For each stage, an absorption factor can be defined by

$$A_n = (L_M/mG_M)_n \quad (26)$$

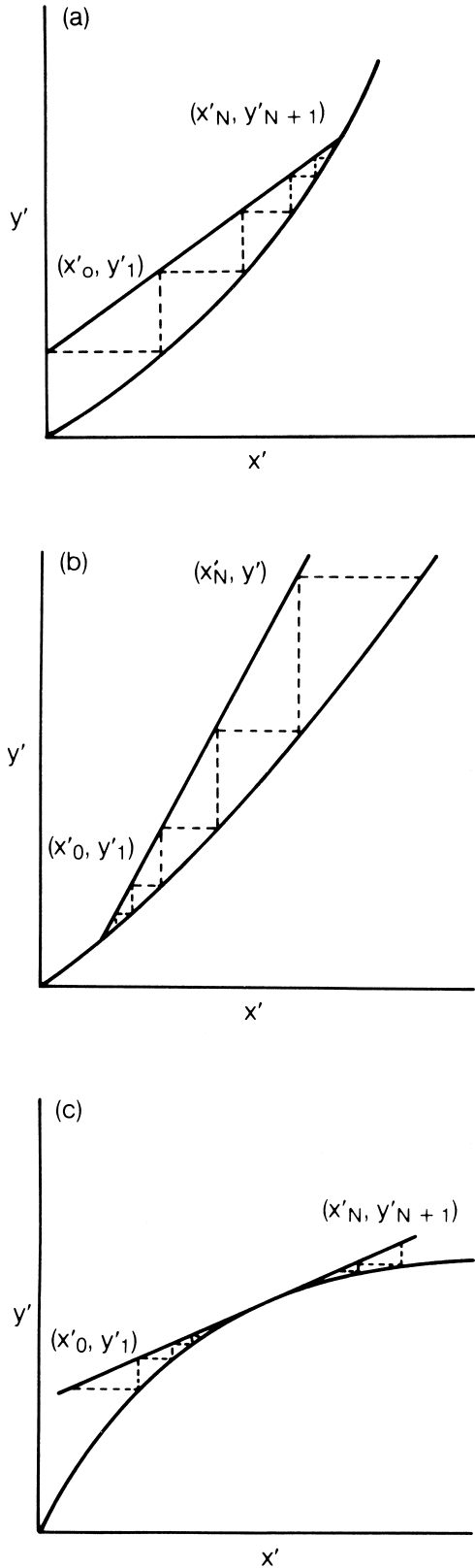


FIGURE 11 Graphic illustrations of minimum solvent rate. (a) Pinch at the bottom, (b) pinch at the top, (c) tangent pinch.

This absorption factor is the ratio of slope of the operating line to that of the equilibrium curve. When the absorption factor is lower than unity, the pinch is located near the bottom of the column (Fig. 11a); when it is higher than unity, the pinch is located near the top of the column (Fig. 11b).

For a dilute gas, and when the equilibrium curve can be approximated by a linear relationship passing through the origin, Eq. (25) is applicable, and an average absorption factor A can be applied to describe the contactor. Under these conditions, an analytical solution of the material balance equation and the equilibrium relationship is possible, giving the Kremser equation:

$$\frac{y_{N+1} - y_1}{y_{N+1} - mx_0} = \frac{A^{N+1} - A}{A^{N+1} - 1} \quad (A \neq 1)$$

$$= N/(N + 1) \quad (A = 1) \quad (27)$$

The left-hand side of Eq. (27) is in principle the ratio of the change of composition of the gas through the contactor to the change that would have occurred had the gas come to equilibrium with the liquid entering the column.

For concentrated gases, the absorption factor varies from stage to stage. In many cases Eq. (27) can be used with an effective average absorption factor and the mole ratio concentration y' :

$$\frac{y'_{N+1} - y'_1}{y'_{N+1} - y'_0} = \frac{A_{\text{ave}}^{N+1} - A_{\text{ave}}}{A_{\text{ave}}^{N+1} - 1} \quad (A_{\text{ave}} \neq 1)$$

$$= \frac{N}{N + 1} \quad (A_{\text{ave}} = 1) \quad (28)$$

The value for A_{ave} is often defined using Eq. (29), with m_{ave} evaluated at the average column temperature:

$$A_{\text{ave}} = L'_M / G'_M m_{\text{ave}} \quad (29)$$

If the absorption is multicomponent, the average equilibrium constant m_{ave} is determined for each of the solute components at the average temperature and pressure of the absorber, and a separate absorption factor A_{ave} is defined for each component. These absorption factors can be used in Eq. (28) to define the absorbed fraction of the component.

Horton and Franklin (1940) used the average absorption factor approach in analyzing a number of absorbers in the petroleum industry. Edmister (1943) extended the Horton and Franklin concept, retaining the Kremser equation form and making use of several empirical factors. He used an effective absorption factor A_e and a modified absorption factor A' , given by

$$A_e = \sqrt{A_N(A_1 + 1) + 0.25} - 0.5 \quad (30a)$$

$$A' = \frac{A_N(A_1 + 1)}{A_N + 1} \quad (30b)$$

Using these definitions, Eq. (28) becomes

$$\frac{y'_{N+1} - y'_1}{y'_{N+1}} = \frac{A_e^{N+1} - A_e}{A_e^{N+1} - 1} \left(1 - \frac{L'_M x'_0}{A' G'_M y'_{N+1}} \right) \quad (30c)$$

Hines and Maddox (1985) found that the Edmister method gives a close approximation to observed or rigorously computed concentration gradients in many multicomponent absorbers.

5. Other Procedures

Graphical procedures such as those described above can also be extended to multicomponent absorption. This subject is discussed in detail by Sherwood (1975).

A method suitable for computer calculations, which carries out tray-by-tray mass, component, and heat balances was first developed by Sujata (1961). In this method, the liquid and vapor flow rates and the temperature profile are assumed and used to calculate an absorption factor for each stage [Eq. (26)]. A component balance is written for each stage in terms of the component flows and absorption factors. The component balances are solved by matrix techniques to give component flows for each stage. Energy balances are then solved to obtain a new temperature profile. The total vapor and liquid flow profiles are found by summing the individual component flows. The calculation is then repeated with the updated temperatures and flows in a trial-and-error manner, until convergence is reached. There are several variations of the above procedure. Some of the popular ones are discussed in Wankat's text. Some rigorous distillation methods have also been extended to absorption.

C. Rate Models

Traditionally, absorbers and strippers were described as stagewise contactors. Krishnamurthy and Taylor developed a new rate (nonequilibrium stage) approach for modeling absorbers and strippers. This approach describes an absorber as a sequence of nonequilibrium stages. Each stage represents a real tray in the case of a tray tower or a section of a continuous contacting device such as a packed column. For each nonequilibrium stage, the mass, component, and energy balance equations for each phase are solved simultaneously, together with the mass and energy transfer rate equations, reaction rate equations, and the interface equilibrium equations. Computation of stage efficiencies is thus avoided altogether and is, in effect, substituted by the rate equations.

Although the rate model can be applied to any separation, it has become most popular in absorption and stripping. Reported case studies demonstrated that, in at least some situations, a rate model can more closely approxi-

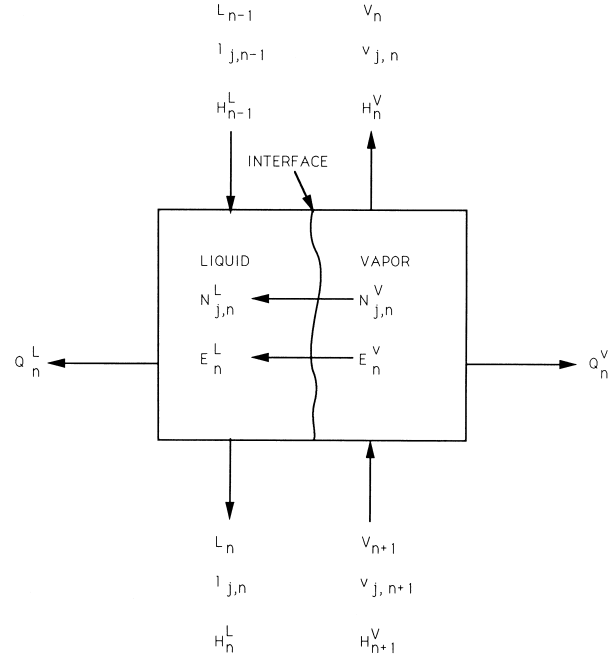


FIGURE 12 Schematic diagram of a nonequilibrium stage n .

mate absorber performance than can an equilibrium stage model. The success of rate models in absorption is largely a result of the difficulty of reliably predicting stage efficiencies in absorbers. The presence of many components, low stage efficiencies, significant heat effects, and chemical reactions are commonly encountered in absorbers and difficult to accommodate in stage efficiency prediction.

Figure 12 is a schematic diagram of a nonequilibrium stage n in an absorber. The equations applying to this stage are described below. A more detailed description is given by Krishnamurthy and Taylor.

Component balances for component j on stage n are given in Eqs. (31a–c) for the vapor phase, the liquid phase, and the interface, respectively:

$$v_{j,n} - v_{j,n+1} + N_{j,n}^V = 0, \quad (31a)$$

$$l_{j,n} - l_{j,n-1} - N_{j,n}^L = 0, \quad (31b)$$

$$N_{j,n}^V = N_{j,n}^L \quad (31c)$$

Energy balances on stage n are given in Eqs. (32a–c) for the vapor phase, the liquid phase, and the interface, respectively:

$$V_n H_n^V - V_{n+1} H_{n+1}^V + Q_n^V + E_n^V = 0 \quad (32a)$$

$$L_n H_n^L - L_{n-1} H_{n-1}^L + Q_n^L - E_n^L = 0 \quad (32b)$$

$$E_n^V = E_n^L \quad (32c)$$

The interface equilibrium is written at the interface.

$$y_{j,n}^I = m_{j,n}^I x_{j,n}^I \quad (33)$$

In Eq. (31c), $N_{j,n}^V$ and $N_{j,n}^L$ are the mass transfer rates. These are calculated from multicomponent mass transfer equations. The equations used take into account the mass transfer coefficients and interfacial areas generated in the specific contactor, reaction rates, heat effects, and any interactions among the above processes.

The above equations, including those describing the mass transfer rates on each stage, are solved simultaneously for all stages. Solution of these nonlinear equations is complex and usually requires a computer. Newton's numerical convergence technique, or a variant of it, is considered to be most effective in solving these equations.

D. Heat Effects in Absorption

When absorption liberates a considerable quantity of heat, and if a large quantity of solute is absorbed, the solution temperature rises. This reduces the solubility of the solute in the liquid, thus counteracting absorption.

The temperature rise can be evaluated from the quantity of heat liberated, which in turn is a function of the change in liquid composition. An equilibrium curve that takes into account the temperature variations through the absorber is shown in Fig. 13, corresponding to a bulk temperature rise from T_2 to T_1 as the bulk liquid composition changes from x_2 to x_1 . The location of the curves depends on which resistance controls, because the equilibrium relationship is obeyed at the interface and not at the bulk.

Work on absorption with large heat effects indicates that the temperature inside an absorber often goes through a

maximum when the solvent is volatile (e.g., ammonia–water absorption).

When solute is absorbed rapidly, the rate of heat liberation is largest near the bottom of the absorber, causing the equilibrium curve to bend upward at the solute-rich end, while remaining relatively unaffected by the heat of solution at the lean end of the absorber. This may sometimes lead to a pinched condition at the rich end of the absorber. When this type of pinching is a concern, it is customary either to provide cooling coils inside the absorber or to divert a liquid stream through an external cooler and then return it to the next lower tray in the column.

Other than the heat of solution, heat effects that may influence absorber performance are solvent vaporization, sensible heat exchange between the gas and the liquid, and loss of sensible heat due to cooling coils or atmospheric cooling. Detailed discussion on heat effects in absorption is presented in the text by [Sherwood et al. \(1975\)](#).

IV. ABSORBER DESIGN

Absorber design is normally carried out in three phases: process design, column sizing, and hydraulic design. In the process design phase, the main system parameters (e.g., solvent selection, operating pressure and temperature, solvent rate, theoretical number of stages, type of contactor) are determined. In the column-sizing phase, the height, diameter, and sizes of the main internals such as downcomers, packings, and tray spacing are determined. Finally, the hydraulic design phase defines all the sizes, dimensions, and layouts of column inlets, outlets, and the multitude of internals used in the column.

A. Process Design

The following steps are followed in column process design:

1. Specification of the separation. A separation is specified by defining column feed flow rate and composition, overhead solute concentration (alternatively, solute recovery), and the concentration of solute (if any) in the lean solvent. If the purpose of absorption is to generate a specific solution, as in acid manufacture, the solution concentration completes the separation specification. For all other purposes, one specifying variable (e.g., rich solvent concentration or solvent flow rate) remains to be specified and is usually set by optimization as outlined below.
2. Selection of solvent and solute recovery process. This was discussed in Section I.
3. Setting the operating pressure. A higher pressure favors the gas solubility and decreases the diameter of

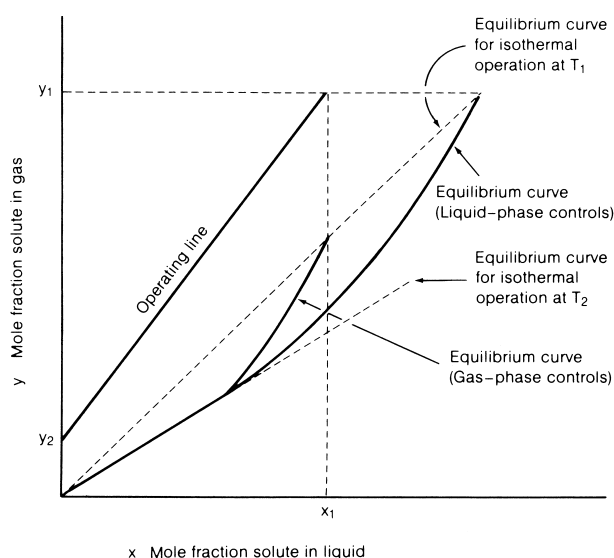


FIGURE 13 Effect of heat liberation on the equilibrium curve.

the contactor. However, the cost of attaining the pressure must be considered. Off-gas scrubbers, for example, process large quantities of gas which is then discharged to the atmosphere; in such a case, the absorber pressure is set at near atmospheric, and the cost of moving the gas through the contactor (due to pressure drop) may govern the decision on operating pressure.

4. Determining solvent circulation rate. If the purpose of absorption is to generate a solution of a specific concentration, the circulation rate is a fixed function of this concentration. For all other purposes, this circulation rate is determined by optimization. As circulation rate is increased, the absorption factor $L_M/(mG_M)$ increases, as does the distance from the operating line to the equilibrium curve. This leads to a shorter and therefore cheaper column. On the other hand, the higher the circulation rate, the greater is the cost of separating the solute from the solvent and the larger is the diameter of the absorber. Many studies have shown that the optimum circulation rate is about 40% greater than the minimum solvent rate.

5. Selection of contactor type. Tray and packed columns are most common; other types are generally used only for special services.

The main factors favoring packed columns are (1) very corrosive applications, where plastic or ceramic packings are favored over trays, which are almost always constructed of metal; (2) low pressure drop requirement, which is easier to achieve with packings than with trays; (3) small-diameter columns, because trays require access for inspection and maintenance; and (4) foaming systems, which are easier to handle in packed towers.

The main factors favoring tray columns are (1) presence of solids (packings have a greater tendency to trap solids and to suffer from the resulting blockage and channeling), (2) very high or very low liquid rates (trays are more suitable to handle these than packings, except for structured packings, which are also capable of handling very low liquid rates), (3) slow reaction rate processes (trays can provide a greater liquid holdup and therefore more residence time), (4) complexities such as cooling coils or intercoolers, which are easier to incorporate into tray columns, and (5) column weight (tray columns are generally lighter and easier to support).

6. The number of theoretical stages, or transfer units, is calculated using a mathematical model of the type described in Section III. At this stage, it is necessary to allow for any heat effects; if these are significant, coiling coils or intercoolers may be required.

B. Column Sizing

In this section, the main types of absorption equipment (packed columns and tray columns) are described, and

the main considerations in their design and sizing are discussed.

1. Packed Columns

A typical arrangement (Fig. 14) consists of a vertical tower containing one or more beds of packings. The descending liquid is distributed over the packing layers, forming liquid films that flow along the surfaces of the particles, thus exposing a large surface area for gas-liquid contact. The solute is transferred from the gas to the liquid across this surface. The type and size of packings may be the same throughout the column or may differ from bed to bed.

The characteristics considered most desirable for good packing performance are a high surface area, a uniform distribution of liquid, and a low resistance to gas flow.

Two types of packings are common: random packings, which are discrete pieces of packings randomly dumped into the column, and structured packings, which are layers of systematically arranged packings, mostly corrugated sheets or wire mesh. Structured packings provide uniform channels for gas and liquid flow, a more even distribution, and greater surface area for the same resistance to gas flow. In general, they tend to lead a more efficient operation but are also more expensive. For absorbers, random packings are more popular, with structured packings being justified only when pressure drop and efficiency demands are unusually high.

Common types of random packings are shown in Fig. 15. The packings shown in Fig. 15a-c have been largely superseded by the packings shown in Fig. 15d-h.

Table II shows several common random packings and compares them on two bases: (1) surface area per unit volume, the larger area providing more opportunity for mass transfer, and (2) packing factor, a measure of throughput capacity and pressure drop, the lower is value, the higher the capacity and the lower the pressure drop. The table shows that as packing size increases, capacity rises while efficiency decreases. The table includes two packings fabricated from plastic (usually polypropylene); this material of construction is resistant to corrosion and is light weight. Plastic packing applications range from sulfuric acid absorbers to off-gas scrubbers and stripping columns.

Table II shows that, as packing size increases, packing capacity rises while packing efficiency decreases. It also shows that both capacity and efficiency are greater for Pall rings and Intalox® saddles than for Raschig rings and Berl saddles.

The data in Table II are approximate, because the geometry of each packing varies slightly from one manufacturer to another. Usually, the type of data shown in this table is provided by each manufacturing company for

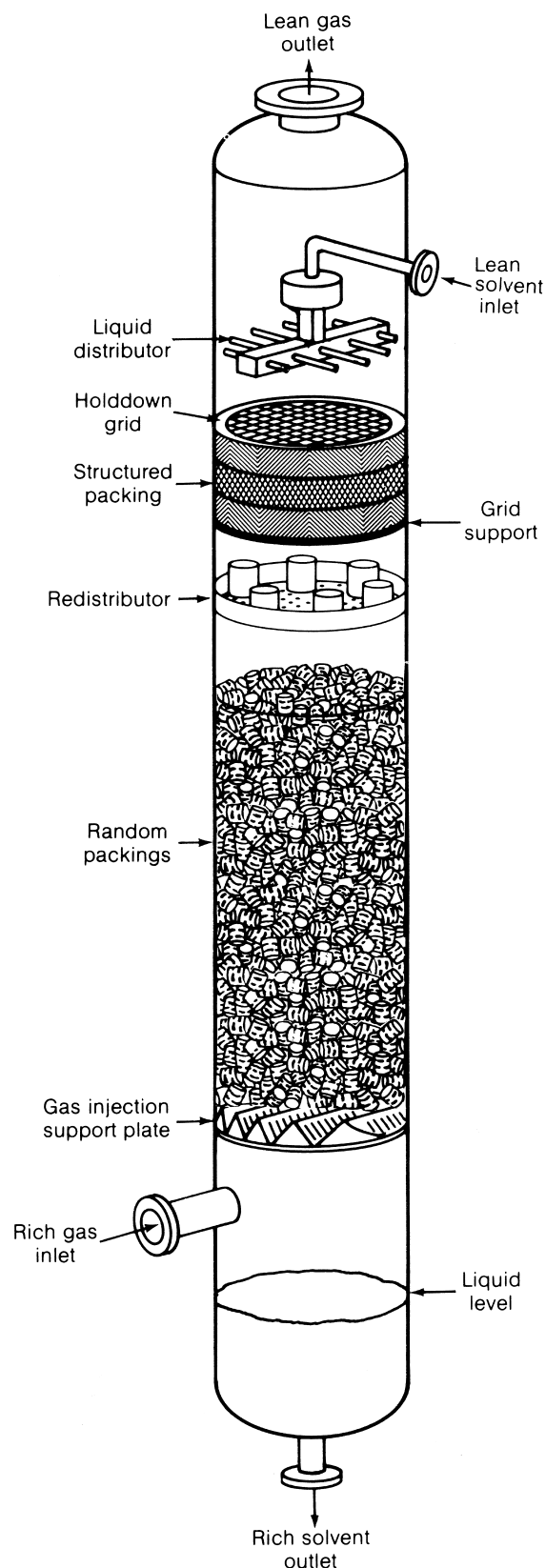


FIGURE 14 Packed column.

its own packings. Perry's text contains a more extensive tabulation of packing factors.

Maximum capacity of a packed bed is usually limited by the onset of flooding. During normal operation, gas flows up while liquid drains freely along the packing surfaces. As gas rate is increased, it begins to interfere with free draining, causing some liquid accumulation in the bed. When this interference is so high that liquid fills the tower, the column is said to be flooded.

The condition of flooding is predicted from generalized charts such as that in Fig. 16. The abscissa shows a scale of a dimensionless term called the flow parameter. This parameter represents a ratio of the kinetic energy of the liquid to the kinetic energy of the gas; thus very low values of the parameter are associated with low-pressure absorbers where the volumetric ratio of gas to liquid may be very high. The ordinate scale shows values of a capacity parameter, generalized through the packing factor (Table II)

Each curve in Fig. 16 represents a constant pressure drop value. Packed absorbers are usually sized to give a pressure drop of 0.25 to 0.50 in. H₂O per foot (200–400 N/m² per meter) of packed depth. Figure 16 is used to determine the column cross sectional area to achieve

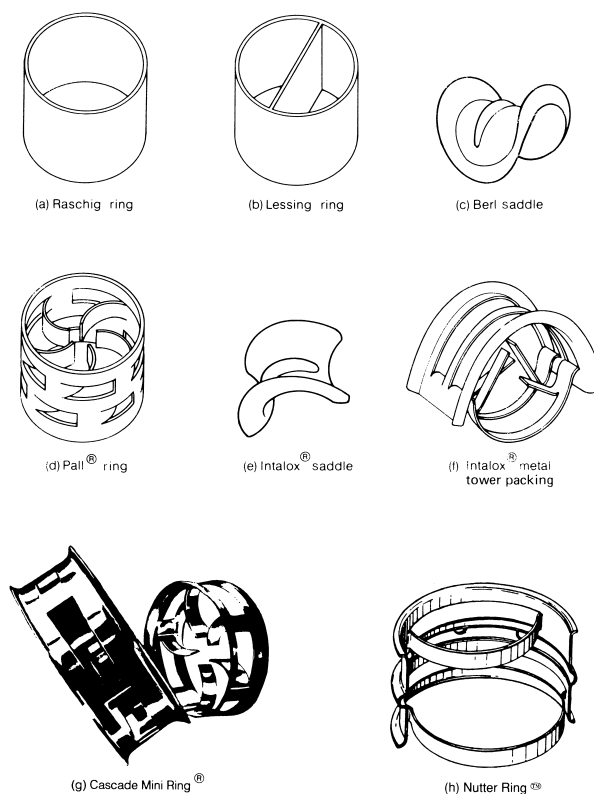


FIGURE 15 Common types of random packings (Parts e and f, courtesy of Norton Co.; part g, courtesy of Glitsch, Inc.; part h, courtesy of Nutter Engineering Corp.)

TABLE II Characteristics of Random Packings^a

Nominal size (mm)	Surface area (m ² /m ³)					Packing factor (m ⁻¹)				
	25	38	50	75	90	25	38	50	75	90
Type										
Raschig ring (metal)	185	130	95	66	—	450	270	187	105	—
Pall ring (metal)	205	130	115	—	92	157	92	66	—	53
Intalox [®] Metal Tower Packing	—	—	—	—	—	135	82	52	43	—
Raschig ring (ceramic)	190	120	92	62	—	510	310	215	120	—
Berl saddle (ceramic)	250	150	105	—	—	360	215	150	—	—
Intalox [®] saddle (ceramic)	255	195	118	92	—	320	170	130	70	—
Intalox [®] saddle (plastic)	206	—	108	88	—	105	—	69	50	—
Pall ring (plastic)	205	130	100	—	85	170	105	82	—	52

^a (From Perry, R. H. (ed.) (1985). "Chemical Engineer's Handbook," 6th ed., McGraw-Hill, New York.)

this pressure drop at the design and liquid loads. Pressure drops of 1.5–1.7 in. H₂O per foot are representative of incipient flooding and values this high are to be avoided.

Packed height is determined from the relationships in Section III. Application of these relationships requires knowledge of the liquid and gas mass transfer coefficients. It is best to obtain these from experimental data on the system if available, but caution is required when extending such data to column design, because mass transfer coefficients depend on packing geometry, liquid and gas distribution, physical properties, and gas and liquid loads, and these may vary from one contactor to another.

In the absence of experimental data, mass transfer coefficients (and hence heights of transfer units) can be estimated by generalized models. A popular and easy to use correlation for random packings is that of Bolles and Fair (1982). The earlier correlations of Onda *et al.* (1968) and Bolles and Fair are also useful for random packings.

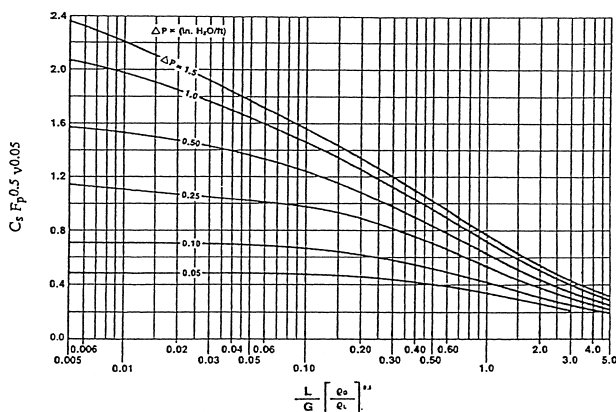


FIGURE 16 Generalized pressure drop correlation of Strigle²¹. C_s = flow parameter = $U_s[\rho_g/(\rho_L - \rho_g)]^{0.5}$, ft/s. F_p = packing factor, ft⁻¹, ν = kinematic viscosity of liquid, centipoises/specific gravity.

For structured packings the correlation of Rocha *et al.* (1996) has been well validated for a number of packings tested in larger equipment. Even if experimental data are available, one must be cautious in applying data taken in small laboratory columns to designs of large commercial contactors.

2. Tray Columns

A typical arrangement (Fig. 17) consists of a vertical tower fitted with horizontal plates or trays, on which liquid and gas are contacted. Each tray is equipped with gas passages, which may be perforations in the tray floor or other devices such as valves or bubble caps that disperse the rising gas into the liquid layer. The liquid layer on the tray is maintained by the outlet weir. Liquid descends from each tray to the tray below via a downcomer.

Liquid enters the column and flows across the top tray, where it contacts the rising gas to form a froth, emulsion, or spray-type dispersion (Fig. 18). It then overflows the weir into the downcomer, which separates gas from the liquid, and carries liquid by gravity to the tray below. The liquid then flows across the next tray, and the process is repeated. Liquid is thus contacted with gas in a stagewise manner.

Two types of trays are most common: sieve trays and valve trays. A sieve tray is a simple perforated plate. Gas issues from the perforations to give a multiorifice effect; liquid is prevented from descending the perforations or "weeping" by the upward motion of the gas. At low gas flow rates, the upward gas motion may be insufficient to prevent weeping.

In valve trays, the perforations are equipped with valve units (Fig. 19). At high gas rates, the gas force opens the valves, thus providing area for gas flow. At low gas rates, there is insufficient force to keep many of the valves open, and these close, preventing the liquid from weeping. Sieve and valve trays show comparable capacity, efficiency, and

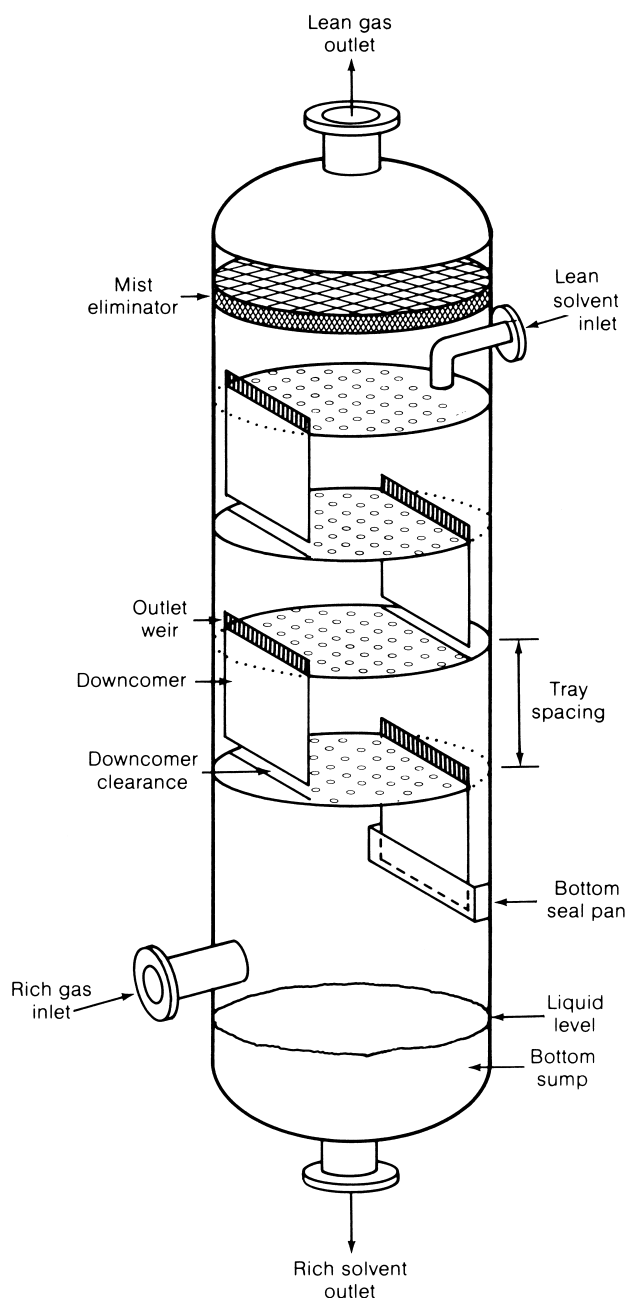


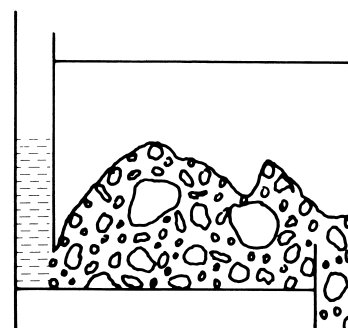
FIGURE 17 Tray column.

other performance characteristics at high gas rates; but valve trays weep less and therefore perform better at low gas rates.

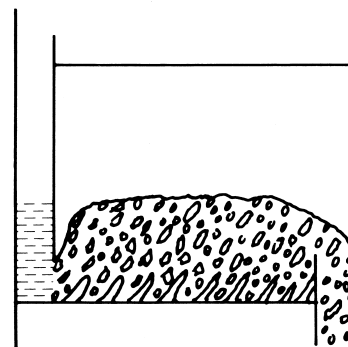
A third type of tray, once commonly employed but currently used only for special applications, is the bubble-cap tray. Its design and operation are discussed by [Bolles \(1963\)](#).

The maximum capacity of a tray column is usually limited by the onset of flooding, which occurs when liquid excessively accumulates inside the column. Flooding is

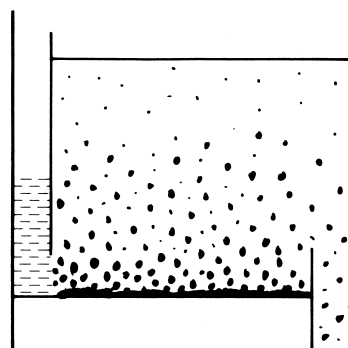
caused by massive liquid carryover from tray to tray (entrainment flood) or when liquid backup in the downcomer reaches the tray above (downcomer backup flood) or when the downcomer is unable to handle the total quantity of descending liquid (downcomer choke flood). At low liquid rates and high gas velocities, entrainment flooding is the most common limit. At high liquid flow rates and low gas velocities (e.g., high pressure operation), downcomer backup and downcomer choke flood are the most common limits.



(a) Froth



(b) Emulsion



(c) Spray

FIGURE 18 Types of dispersion on an absorption tray.

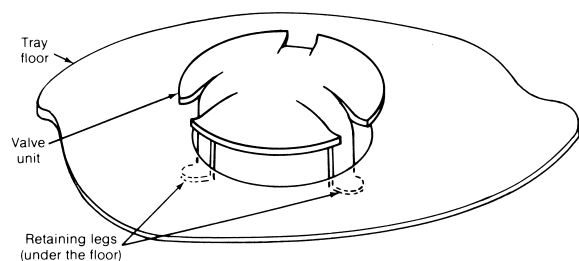


FIGURE 19 Flexitray valve unit (courtesy of Koch Engineering Company, Inc.).

Entrainment flooding is predicted by an updated version of the Souders and Brown correlation. The most popular is Fair's (1961) correlation (Fig. 20), which is suitable for sieve, valve, and bubble-cap trays. Fair's correlation gives the maximum gas velocity as a function of the flow parameter $(L/G)\sqrt{(\rho_G/\rho_L)}$, tray spacing, physical properties, and fractional hole area.

Downcomer backup flooding occurs when the backup of aerated liquid in the downcomer exceeds the available tray spacing. Downcomer backup can be calculated by adding the clear liquid height on the tray, the liquid backup caused by the tray pressure drop, and the liquid backup caused by the friction loss at the downcomer outlet. The downcomer backup is then divided by an aeration factor to give the aerated liquid backup.

To avoid downcomer choke flooding, downcomers are sized to give a liquid residence time of not less than 3–7 sec, depending on the tendency of the liquid to form a stable foam.

Tray area is usually determined from an entrainment flooding correlation. Trays are normally designed to operate at 80 to 85% of flood at the maximum expected throughput. Downcomer area is usually determined from the downcomer choke criteria. The design is then checked to ensure that downcomer backup flood does not occur.

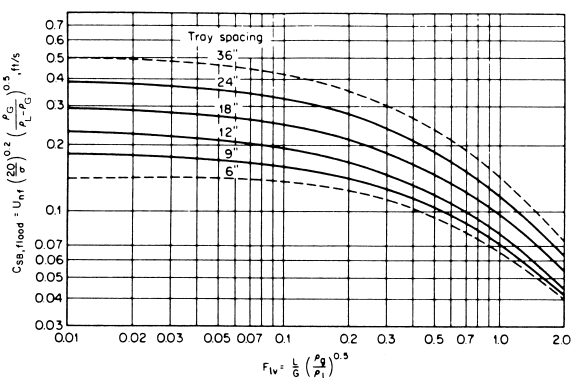


FIGURE 20 Entrainment flooding correlation for trays. (From Fair, J. R. (1961). *Petrol Chem. Engineer* Sept., p. 45; reproduced by permission of Petroleum Engineer International, Dallas, Texas.)

The number of trays is determined by dividing the theoretical number of stages, which is obtained from the relationships in Section III, by the appropriate tray efficiency. It is best to use experimental efficiency data for the system when available, but caution is required when extending such data to column design, because tray efficiency depends on tray geometry, liquid and gas loads, and physical properties, and these may vary from one contactor to another. In the absence of data, absorption efficiency can be estimated using O'Connell's empirical correlation. This correlation should not be used outside its intended range of application.

During the column-sizing phase, a preliminary tray layout is prepared by setting the following:

1. Tray spacing. Eighteen to 24 in. (450–600 mm) is considered optimum, but smaller or larger values are not uncommon; for example, smaller values are used if total column height is restricted. A lower tray spacing leads to a shorter column at the expense of a greater diameter.
2. Number of liquid passes. At high liquid flow rates, the liquid may be split into two or more paths. This reduces the effective liquid loads, leading to a higher capacity at the expense of a shorter flow path and therefore lower efficiency.
3. Fractional hole area (sieve trays). Eight to 10% is generally considered optimum. Higher area may enhance capacity at the expense of more weeping at low gas flow rates.
4. Weir height. This parameter sets the level of liquid on the tray in the froth and emulsion regimes (Fig. 17a,b). The higher the level, the better is the contact and the efficiency at the expense of a greater liquid backup in the downcomer. Typical absorption weir heights are 2–3 in. (50–75 mm).
5. Downcomer sloping. Sloped downcomers are often used to permit a greater perforated tray area while maintaining a high downcomer entrance area, needed to prevent downcomer choke.
6. Downcomer clearance. A high clearance increases downcomer capacity at the expense of increasing the tendency of the downcomer to pass vapor. A common design practice is to set the clearance to 0.25 to 0.5 in. (6–13 mm) less than the weir height.

3. Other Contactors

Other contactor types used for absorption include the following:

Spray columns. These are columns fitted with rows of sprays located at different heights. Gas rises vertically, and liquid is sprayed downward at each of these rows. Mass transfer is usually poor because of low gas and liquid

residence times and because of extensive gas backmixing. Their application is limited to easy absorption duties (one or two theoretical stages), usually in systems where the controlling resistance to mass transfer is in the gas phase. Column capacity is usually limited by liquid droplet entrainment from the top.

Spray absorbers are advantageous where low pressure drops are critical and where the gas may contain some solids, such as in the absorption of SO_2 from coal-fired boiler exhaust gases.

Falling-film absorbers. These are usually vertical heat exchangers with the cooling medium in the shell and the absorption taking place in the tubes. The solvent flows downward, while the gas may enter either at the bottom (countercurrent flow) or at the top (cocurrent flow).

Mass transfer in falling-film absorbers is strongly dependent on the gas velocity in the tubes, the liquid and gas distribution, and the tube surface conditions. The maximum capacity of falling-film absorbers is normally restricted either by flooding or by pressure drop. Another important limit in these absorbers is film breakup. If heat flux is excessive, dry areas may form at the tube wall and reduce mass transfer.

Falling-film absorbers make continuous heat removal possible and are therefore extensively used in applications where the heat released during absorption is high, such as in the absorption of hydrogen chloride to form hydrochloric acid.

Stirred tanks. These are mechanically stirred vessels, which are advantageous when absorption is accompanied by a slow liquid-phase chemical reaction. As discussed earlier (Section II), this application is considered a chemical reactor rather than an absorber. Stirred tanks provide high liquid residence times but are limited to low gas flow rates.

Bubble columns. These are columns full of liquid into which gas is introduced by a perforated pipe or a sparger. Bubble columns are used for applications similar to stirred tanks, but their contact efficiency is lower.

Venturi scrubbers. In a venturi scrubber, a liquid jet issues from a nozzle. The jet induces cocurrent gas flow into the throat of the jet. Mass transfer takes place between the gas and the atomized liquid downstream of the nozzle. Mass transfer is usually poor and depends on the throat velocity or pressure drop, the liquid/gas ratio, and the liquid atomization pattern. Because of the cocurrent nature of contacting, the maximum solute removal does not exceed a single theoretical stage. Venturi scrubbers are used primarily for separation of fine particulate matter or

fine liquid mist from a gas stream. They are often also used for simultaneously absorbing certain components from the gas stream, but because of their poor mass transfer are effective only when these components are highly soluble in the liquid. Common applications are scrubbing incinerator fumes and sulfuric and phosphoric acid mists.

Wet scrubbers. These are devices in which a liquid spray contacts a gas stream, primarily for the purpose of removing fine solid particles or liquid mists from the gas. In this process, the liquid spray simultaneously absorbs soluble components from the gas. The sprays are generated by a variety of mechanical devices.

C. Hydraulic Design

This design phase determines the types, dimensions, location, and orientation of the multitude of internals used in absorption columns. It usually leads to refinements to the column design and sizing and, most important, is critical for ensuring trouble-free operation.

1. Packed Columns

The most important aspects of packed-column internals and their design are outlined in the following paragraphs.

Packed-tower efficiency and turndown are strongly dependent on the quality of initial *liquid distribution*. Uneven distribution may cause local variations in the liquid/gas ratio, localized pinch conditions, and reduced vapor-liquid contact. [Figure 14](#) shows two common liquid distributor types, the ladder type (shown as the top distributor) and the orifice type (shown as the redistributor). The ladder type is a horizontal header of pipes, which are perforated on the underside. The orifice type is a flat perforated plate equipped with round or rectangular risers for gas passage. Other common types of distributors are a header equipped with spray nozzles (spray distributor) and a header of horizontal channels, with V notches cut in the vertical walls of the channels (notched-trough distributor).

Ladder and spray distributors rely on pressure for their action. They provide a large gas flow area but a somewhat limited liquid flow area; they are light and cheap but are sensitive to corrosion, erosion, and to a certain extent plugging. They are most suitable for high gas/liquid ratio applications.

Orifice and notched-trough distributors rely on gravity for their action. They provide a large liquid flow area; the notched-trough distributor also provides a large gas flow area. They are more robust and expensive than pressure distributors and are sensitive to levelness. The orifice distributor is most sensitive to plugging, while the notched-trough is the least sensitive to plugging, corrosion,

or erosion. The orifice distributor has the potential to generate a distribution pattern superior to most others, but its application is often restricted to clean fluids where the gas/liquid ratio is not high. The notched-trough distributor is often considered the most reliable distributor, although the quality of distribution may be somewhat inferior than that of the orifice or ladder distributors.

Liquid redistributors are installed at frequent intervals in a packed column to remix the liquid, thus counteracting the propagation of maldistribution effects and the natural tendency of liquid to migrate toward the wall. A common design practice is to redistribute the liquid every 20 ft (6–7 m).

Redistributor design is similar to gravity distributor design. The orifice type is most popular (Fig. 14). A notched-through type requires a liquid collection device above it to feed the liquid onto the distributor. Often, the gas risers are equipped with caps to prevent liquid from dropping through the gas spaces.

Liquid collectors are installed when liquid must be collected for redistribution or drawoff (e.g., for external cooling). The common device used is a chimney tray, which is similar to an orifice redistributor, but without perforations. Another common device is the Chevron-type collector, which is a series of Chevron blades, with liquid being collected at the bottom of the blades.

Packing supports have to support the packed bed physically, while incorporating a large free area to permit free passage of gas and liquid. Grid supports are common, especially in nonmetallic applications. Gas injection supports (Fig. 14) are usually preferred; these provide separate passages for the gas and liquid and a large free area.

Holddown plates and bed limiters are grids or wire screens with openings small enough to prevent migration of packing particles. They prevent bed fluidization, which may cause breakage of ceramic and carbon packings or entrainment of metal or plastic packings with the gas.

2. Tray Columns

The most important features of tray column internals and their designs are outlined in the following paragraphs.

Liquid inlets. Liquid enters the top tray via a hole in the column shell, often discharging against a vertical baffle or weir, or via a short, down-bending pipe (Fig. 17), or via a distributor. Restriction, excessive liquid velocities, and interference with tray action must be avoided, as these may lead to excessive entrainment, premature flooding, and even structural damage. Disperser units (e.g., perforations, valves) must be absent in the liquid entrance area (Fig. 17) or excessive weeping may result.

Gas inlets. Gas must enter above the bottom liquid level or, if bubbled through the liquid, through a well-designed sparger. Commonly, no sparger is used; in such cases, the feed nozzle should be located at least 12 in. (0.3 m) above the liquid level. Impingement on the liquid level, seal pan overflow, and instrument nozzles must be avoided. Failure to follow these guidelines may result in premature flooding, excessive entrainment, and in some cases mechanical damage to the trays.

Bottom liquid outlets. Sufficient residence time must be provided in the bottom of the column to separate any entrained gas from the leaving liquid. Gas in the bottom outlet may also result from vortexing or from forthing caused by liquid dropping from the bottom tray (a “waterfall pool” effect). Vortex breakers are commonly used, and liquid-drop height is often restricted. Inadequate gas separation may lead to bottom pump cavitation or vapor choking the outlet line.

Intermediate liquid outlets. Liquid may be withdrawn using a chimney tray or from a downcomer. A chimney tray is a flat, unperforated plate with vapor risers. It permits total withdrawal of liquid; a downcomer drawoff permits only partial withdrawal because some weeping occurs through the tray. A downcomer drawoff may contain some entrained gas, which must be separated downstream or allowed for in downstream equipment design.

Gas outlets. Sufficient liquid disentrainment from the overhead gas is usually required. This may be achieved by providing sufficient vertical height above the top tray, installation of mist eliminators, or providing external knockout facilities downstream of the column.

Tray layout. The preliminary tray and downcomer layout is prepared in the column-sizing phase and refined during the hydraulic design phase. In addition to the parameters previously set, such parameters as hole diameter or the type of valve unit are determined.

Smaller hole diameters usually enhance efficiency and capacity but are also more sensitive to corrosion and plugging. Holes smaller than $\frac{3}{16}$ in. (5 mm) are uncommon because they require an expensive manufacturing technique. Half-inch (13-mm) holes are common when corrosion or plugging is expected.

The best type of valve unit depends on the corrosive and fouling tendencies of the service, as some valve units tend to pop out of their seats in corrosive services, while others tend to stick to their seats in fouling services.

Other parameters such as level tolerance, tray supports, drainage, weir shape and type are also determined in this phase.

Downcomers layout. Usually, segmental downcomers are used, in which the downcomer area extends from the weir to the column wall (Fig. 17), but other designs are not uncommon. The design must consider downcomer hydraulics as well as mechanical and structural factors.

The need for positively sealing the downcomer is determined in this phase. This could be achieved by installing an inlet weir, which is a weir installed at the tray inlet to keep the downcomer outlet immersed in liquid. A similar device, which extends below the tray floor, is a seal pan (Fig. 17). Both devices provide positive assurance against vapor rising up the downcomer, but they may also trap solids and dirt and cause blockage. A seal pan must always be used in the downcomer from the bottom tray; otherwise there is nothing to prevent vapor from rising up the bottom downcomer.

NOMENCLATURE

A	Component A	F_p	Packing factor, ft^{-1} (m^{-1})
A	Absorption factor, $L_M/(mG_M)$, dimensionless	G	Gas flow rate (Fig. 16 only), $\text{lb}/(\text{s ft}^2)$ ($\text{kg}/(\text{s m}^2)$)
a	Effective interfacial mass transfer area per unit volume, ft^2/ft^3 (m^2/m^3)	G	Gas flow rate, lb/h (kg/h)
A'	Modified absorption factor, given by Eq. (31b)	g_c	Conversion factor, 32.2 ($\text{lb ft}/(\text{lbf s}^2)$) ($1.0(\text{kg m})/(\text{N s}^2)$)
A_e	Effective absorption factor, given by Eq. (31a)	G_M	Molar gas-phase mass velocity, $\text{lb mol}/(\text{h ft}^2)$ [$\text{kmol}/(\text{s m}^2)$]
B	Component B	G'_M	Molar gas-phase mass velocity of rich gas, $\text{lb mol}/(\text{h ft}^2)$ [$\text{kmol}/(\text{s m}^2)$]
b	Number of moles of component B reacting with 1 mole of component A	H	Enthalpy, $\text{Btu}/\text{lb mole}$ (kJ/kmol) (Fig. 12 and Eq. (33) only)
C	Component C	H	Henry's Law constant, atm (kPa)
c	Number of moles of component C produced when 1 mole of component A reacts with b moles of component B	h	Height parameter for packed towers, ft (m)
C_A	Concentration of reactant A in the liquid, $\text{lb mole}/\text{ft}^3$ ($\text{kg mole}/\text{m}^3$)	H_a	Hatta number, defined by Eq. (12), dimensionless
C_B	Concentration of reactant B in the liquid, $\text{lb mole}/\text{ft}^3$ ($\text{kg mole}/\text{m}^3$)	H_G	Height of a transfer unit based on gas-phase resistance, ft (m)
C_{B_0}	Concentration of reactant B in the bulk liquid, $\text{lb mole}/\text{ft}^3$ ($\text{kg mole}/\text{m}^3$)	H_L	Height of a transfer unit based on liquid-phase resistance, ft (m)
C_{SB}	Flooding capacity parameter, given in Fig. 20, ft/s (m/s)	H_{OG}	Height of an overall gas-phase mass-transfer unit, ft (m)
D_A	Diffusion coefficient of component A in the liquid phase, ft^2/h (m^2/s)	h_T	Contact height, ft (m)
D_B	Diffusion coefficient of component B in the liquid phase, ft^2/h (m^2/s)	k_2	Second order reaction rate constant, $\text{ft}^3/(\text{h lb mol})$ [$\text{m}^3/(\text{s kmol})$]
E	Energy transfer rate across interface, Btu/h (kJ/s)	k_G	Gas-phase mass-transfer coefficient for dilute systems, $\text{lb mol}/(\text{h ft}^2 \text{ mole fraction solute})$ ($\text{kmol}/(\text{s m}^2 \text{ mole fraction solute})$)
F_{IV}	Flow parameter, $(L/G) \sqrt{\rho_G/\rho_L}$, dimensionless	k'_G	Gas-phase mass-transfer coefficient for concentrated systems, same units as k_G
		k''_G	Gas-phase mass transfer coefficient for multicomponent systems, same units as k_G
		k_L	Liquid-phase mass-transfer coefficient for dilute systems, same units as k_G
		k'_L	Liquid-phase mass-transfer coefficient for concentrated systems, same units as k_G
		k''_L	Liquid-phase mass transfer coefficient for multicomponent systems, same units as k_G
		k_L^0	Liquid-phase mass-transfer coefficient for pure physical absorption (no reaction), same units as k_G
		K_{OG}	Overall gas-phase mass-transfer coefficient for dilute systems, same units as k_G
		K'_{OG}	Overall gas-phase mass-transfer coefficient for concentrated systems, same units as k_G
		K''_{OG}	Overall gas-phase mass transfer coefficient for multicomponent systems, same as units as k_G
		L	Liquid flow rate (Fig. 16 only), $\text{lb}/(\text{s ft}^2)$ ($\text{kg}/(\text{s m}^2)$)

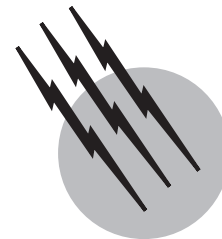
L	Liquid flow rate, lb/h (kg/h)	x_{fm}	Film factor, given by Eq. (8b)
L	Liquid flow rate, lb mole/h (kmol/s) (Fig. 12 and Eq. 33 only)	y	Mole fraction solute (in bulk-gas phase, unless otherwise subscripted)
l	Liquid component flow rate, lb mole/h (kmol/s)	y'	Mole solute in gas per mole of rich gas entering the absorber
L_M	Molar liquid-phase mass velocity, lb mol/(h ft ²) [kmol/(s m ²)]	y_A	Mole fraction solute A (in bulk-gas phase, unless otherwise subscripted)
L'_M	Molar solute-free solvent mass velocity, lb mol/(h ft ²) [kmol/(s m ²)]	y^*	Mole fraction solute in bulk-gas in equilibrium with solute concentration in bulk-liquid
m	Slope of equilibrium curve = dy^*/dx , dimensionless	y_A^*	Mole fraction solute in bulk-gas in equilibrium with solute concentration in bulk-liquid
N	Number of stages in a stagewise contactor	y_{BM}	Logarithmic-mean inert-gas concentration between bulk-gas and interface, defined by Eq. (6a)
N	Mass transfer rate across interface, lb mole/h (kmol/s) (Fig. 12 and Eq. (32) only)	y_{BM}^*	Logarithmic-mean inert-gas concentration between bulk-gas and value in equilibrium with bulk-liquid
N_A	Molar flow rate of solute A per unit interfacial area, lb mol/(h ft ²) [kmol/(s m ²)]	y_{fm}	Film factor, given by Eq. (8b)
$N_B, N_C, \dots$	As N_A , but with respect to solute B, C, ...	y_{fm}^*	Film factor, given by Eq. (8d)
N_G	Number of gas-phase mass-transfer units, dimensionless	δ	Film thickness, ft (m)
N_L	Number of liquid-phase mass-transfer units, dimensionless	μ	Liquid viscosity, cP [kg/(s m)]
N_{OG}	Number of overall gas-phase mass-transfer units, dimensionless	ρ	Density, lb/ft ³ (kg/m ³)
P	Pressure, atm (kPa)	σ	Surface tension, dyn/cm (N/m)
p	Solute partial pressure in bulk gas, atm (kPa)	ϕ	Ratio k_L/k_L^0 , reaction enhancement factor, dimensionless
Q	Heat removal rate, Btu/h (kJ/s)	ϕ_∞	Ratio k_L/k_L^0 when $H_a = \infty$, dimensionless
r_A	Volumetric reaction rate of component A, lb mol/(h ft ³) [kmol/(s m ³)]	ψ	Ratio of water to liquid density, dimensionless
r_B	Volumetric reaction rate of component B, lb mol/(h ft ³) [kmol/(s m ³)]	Subscripts	
T	Temperature, °F (°C)	0	Liquid inlet to stage contactor
t	Parameter defined by Eq. (8d), indicating degree of counter-diffusion	1	Column bottom (differential contactor)
U_{nf}	Vapor velocity, based on tray area less the area at the bottom of the downcomer, at the flood point, ft/s (m/s)	1	Stage 1 (Top stage in a stagewise contactor)
U_s	Superficial vapor velocity, ft/s	2	Column top (differential contactor)
V	Gas flow rate, lb mole/h (kmol/s)	2	Stage 2 (Stagewise contactor)
v	Gas component flow rate, lb mole/h (kmol/s)	A	Component A
x	Mole fraction solute (in bulk-liquid phase unless otherwise subscripted)	ave	Average for the column
x'	Mole solute in liquid per mole of solute-free solvent entering absorber	B	Component B
x_A	Mole fraction solute A (in bulk-liquid phase, unless otherwise subscripted)	C	Component C
x_A^*	Mole fraction solute in bulk-liquid inequilibrium with solute concentration in bulk-gas	G	Gas phase
x_{BM}	Logarithmic-mean inert-solvent concentration between bulk liquid and interface, given by Eq. (6b)	i	Interface
		j	Component j
		L	Liquid phase
		N	Stage N (bottom stage in a stagewise contactor)
		n	Stage n
		Superscripts	
		I	At the interface
		L	Liquid
		v	Vapor (or gas)

SEE ALSO THE FOLLOWING ARTICLES

ADSORPTION (CHEMICAL ENGINEERING) • CHEMICAL THERMODYNAMICS • ELECTROLYTE SOLUTIONS, THERMODYNAMICS • KINETICS (CHEMISTRY) • NONELECTROLYTE SOLUTIONS, THERMODYNAMICS

BIBLIOGRAPHY

- Bolles, W. L. (1963). Chap. 14. In "Design of Equilibrium Stage Processes" (B. D. Smith, ed.), McGraw-Hill, New York.
- Bolles, W. L., and Fair, J. R. (1982). *Chem. Eng.* **89** (14), 109–116.
- Bravo, J. L., and Fair, J. R. *Ind. Eng. Chem. Proc. Des. Devel.* **21**, 162–179.
- Chan, H., and Fair, J. R. (1984). *Ind. Eng. Chem. Proc. Des. Devel.* **23**, 814–819.
- Danckwerts, P. V. (1970). "Gas-Liquid Reactions," McGraw-Hill, New York.
- Edmister, W. C. (1943). *Ind. Eng. Chem.* **35**, 837–839.
- Fair, J. R. (1961). *Petro/Chem. Eng.* **33** (10), 45–52.
- Fair, J. R. (1997). Gas absorption and gas-liquid system design, Chap. 14. In "Perry's Chemical Engineers' Handbook," 7th ed. (Perry, R. H., and Green, D. eds.), McGraw-Hill, New York.
- Fair, J. R., Bolles, W. L., and Null, H. R. (1983). *Ind. Eng. Chem. Proc. Design Devel.* **22**, 53–58.
- Fredenslund, A., Gmehling, J., and Rasmussen (1977). "Vapor-Liquid Equilibria Using UNIFAC," Elsevier, Amsterdam.
- Hildebrand, J. H., Prausnitz, J. M., and Scott, R. L. (1970). "Regular and Related Solutions," Van Nostrand-Reinhold, New York.
- Hines, A. L., and Maddox, R. N. (1985). "Mass Transfer," Prentice-Hall, Englewood Cliffs, NJ.
- Hobler, T. (1966). "Mass Transfer and Absorbers" (Engl. Ed.), Pergamon, Oxford, U.K.
- Horton, G., and Franklin, W. B. (1940). *Ind. Eng. Chem.* **32**, 1384.
- Hwang, Y.-L., Keller, G. E., and Olson, J. D. (1992). *Ind. Eng. Chem. Res.* **31**, 1759.
- O'Connell, H. E. (1946). *Trans. AIChE* **42**, 741–755.
- Onda, K., Takeuchi, H., and Okumoto, Y. (1968) *J. Chem. Eng. Japan* **1** (1), 56.
- Prausnitz, J. M., and Shair, F. M. (1961). *AIChE J.* **7**, 682–687.
- Reid, R. C., Prausnitz, J. M., and Poling, B. E. (1987). "The Properties of Gases and Liquids," 4th ed., McGraw-Hill, New York.
- Rocha, J. A., Bravo, J. L., and Fair, J. R. (1996). *Ind. Eng. Chem. Res.* **35**, 1660–1667.
- Rousseau, R. W., and Staton, J. S. (1988). *Chem. Eng.* **95**, 91–95.
- Sherwood, T. K., Pigford, R. L., and Wilke, C. R. (1975). "Mass Transfer," 3rd Ed., McGraw-Hill, New York.
- Strigle, R. F. (1994). "Packed Tower Design and Applications," 2nd Ed., Gulf Publ. Co., Houston.
- Sujata, A. D. (1961). *Hydrocarbon Proc. Petrol. Refiner* **40** (12), 137.
- Van Krevelen, D. W., and Hoftijzer, P. J. (1948). *Rec. Trav. Chim.* **67**, 563.
- Wilke, C. R. (1950). *Chem. Eng. Prog.* **46**, 95.



Adsorption (Chemical Engineering)

Douglas M. Ruthven

University of Maine

- I. Forces of Adsorption
- II. General Applications
- III. Microporous Adsorbents
- IV. Adsorption Equilibrium
- V. Adsorption Kinetics
- VI. Adsorption Column Dynamics
- VII. Cyclic Batch Adsorption Processes
- VIII. Chromatographic Processes
- IX. Continuous Countercurrent Processes

GLOSSARY

Breakthrough curve Plot showing variation of outlet concentration of one (or more) of the adsorbable species with time.

Carbon molecular sieve Microporous carbon adsorbent that has very small micropores (typically $\sim 5.0\text{-}\text{\AA}$ diameter) with a very narrow distribution of pore size.

Extract Product stream containing the more strongly adsorbed species.

HETP Height equivalent to a theoretical plate. A measure of the combined effects of axial mixing and finite mass transfer resistance in causing deviations from ideal (equilibrium) behavior in a chromatographic column or in a countercurrent contact system. The definitions of HETP in these two cases are somewhat different,

reflecting the difference in the flow pattern, but there is a well-defined relationship between the two quantities.

Knudsen diffusion Mechanism of diffusion, dominant in smaller macropores at relatively low pressures, when collisions between diffusing molecules and pore walls occur more frequently than collisions between the molecules themselves.

Langmuir isotherm or model Simple mathematical representation of a favorable (type I) isotherm defined by Eq. (2) for a single component and Eq. (4) for a binary mixture. The separation factor for a Langmuir system is independent of concentration. This makes the expression particularly useful for modeling adsorption column dynamics in multicomponent systems.

LUB Length of unused bed. See Eq. (25) and Fig. 9 for a precise definition.

Macropore diffusion Diffusion in “macropores”—pores that are large compared with the molecular diameter. Several different mechanisms contribute to macropore diffusion, notably ordinary molecular diffusion in larger macropores at higher pressures or in liquids and Knudsen diffusion in smaller macropores at low pressures. Also referred to as intraparticle diffusion.

Mass transfer zone Region in an adsorption column where, at a given time, the concentration of one of the adsorbable species varies with distance along the column.

Micropore diffusion Diffusion within the small micropores of the adsorbent which are of a size comparable with the molecular diameter of the sorbate. Under these conditions the diffusing molecule never escapes from the force field of the solid surface and steric hindrance is important. For zeolites the terms *micropore diffusion* and *intracrystalline diffusion* are synonymous.

Raffinate Product stream containing the less strongly adsorbed species.

Selectivity Difference in the affinity of the adsorbent for two components. Measured quantitatively by the “separation factor,” *q.v.*

Separation factor Defined according to Eq. (5) in analogy with relative volatility; provides a quantitative measure of selectivity.

Zeolite Microporous crystalline aluminosilicate. In this article the term is used in its broad sense to include microporous crystalline silica and aluminophosphates as well as true zeolites.

ADSORPTION is the adhesion or retention of a thin layer of molecules of a gas or liquid mixture brought into contact with a solid surface resulting from the force field at the surface. Because the surface may exhibit different affinities for the various components of a fluid, the composition of the adsorbed layer generally differs from that of the bulk fluid. This phenomenon offers a straightforward means of purification (removal of undesirable components from a fluid mixture) as well as a potentially useful method of bulk separation (separation of a mixture into two or more streams of enhanced value).

I. FORCES OF ADSORPTION

Adsorption is conveniently considered as either “physical adsorption” or “chemisorption,” depending on the nature and strength of the surface forces. Chemisorption can be considered as the formation of a chemical bond between the sorbate and the solid surface. Such interactions are strong, highly specific, and often not easily reversible.

Chemisorption systems are sometimes used for removing trace concentrations of contaminants, but the difficulty of regeneration makes such systems unsuitable for most process applications so most adsorption processes depend on physical adsorption. The forces of physical adsorption are weaker than the forces of chemisorption so the heats of physical adsorption are lower and the adsorbent is more easily regenerated. Several different types of force are involved. For nonpolar systems the major contribution is generally from dispersion–repulsion (van der Waals) forces, which are a fundamental property of all matter. When the surface is polar, depending on the nature of the sorbate molecule, there may also be important contributions from polarization, dipole, and quadrupole interactions. Selective adsorption of a polar species such as water or a quadrupolar species such as CO₂ from a mixture with other nonpolar species can therefore be accomplished by using a polar adsorbent. Indeed, adjustment of surface polarity is one of the main ways of tailoring adsorbent selectivity.

The strength of the van der Waals interaction is directly related to the polarizability of the sorbate which depends, in turn, on the molecular weight. The affinity sequence for nonpolar sorbates therefore generally correlates approximately with the sequence of molecular weights.

Water is a small and highly polar molecule. It is therefore adsorbed strongly on a polar surface, and such adsorbents are therefore commonly called “hydrophilic.” By contrast, water is adsorbed only weakly on a nonpolar surface so such adsorbents are called “hydrophobic.” However, this is something of a misnomer since water is not actually repelled by a nonpolar surface.

II. GENERAL APPLICATIONS

A wide range of adsorption processes have been developed, and such processes are in common industrial use, particularly in the petroleum and petrochemical industries.

The traditional application of adsorption in the process industries has been as a means of removing trace impurities from gas or liquid streams. Examples include the removal of H₂S from hydrocarbon streams before processing, the drying and removal of CO₂ from natural gas, and the removal of organic compounds from waste water. In these examples the adsorbed component has little value and is generally not recovered. Such processes are generally referred to as *purification processes*, as distinct from *bulk separations*, in which a mixture is separated into two (or more) streams, each enriched in a valuable component, which is recovered. The application of adsorption to bulk separations is a more recent development that was stimulated to a significant extent by the rapid

escalation of energy prices during the 1970s. The traditional method of bulk separation is distillation, and although distillation has the advantages of wide applicability and proven technology, it suffers from the disadvantage of very poor energy efficiency, particularly when the difference in volatility of the components to be separated is small. With increasing energy costs the balance of economic advantage for such separations has shifted toward alternative technologies, such as adsorption, that generally involve a higher capital outlay but offer the advantage of greater energy efficiency and therefore lower operating costs. Examples of large-scale bulk separation processes that are commonly accomplished by adsorption include the separation of xylene isomers (liquid phase), the separation of linear and branched paraffins (gas phase or liquid phase), and the separation of olefins from paraffins (gas phase or liquid phase). Similar adsorption separation processes have also been developed for a number of important carbohydrate separations (e.g., fructose–glucose) that cannot easily be accomplished by more traditional methods.

The primary requirement for an economic adsorption separation process is an adsorbent with sufficient selectivity, capacity, and life. Adsorption selectivity may depend either on a difference in adsorption equilibrium or, less commonly, on a difference in kinetics. Kinetic selectivity is generally possible only with microporous adsorbents such as zeolites or carbon molecular sieves. One can consider processes such as the separation of linear from branched hydrocarbons on a 5A zeolite sieve to be an extreme example of a kinetic separation. The critical molecular diameter of a branched or cyclic hydrocarbon is too large to allow penetration of the 5A zeolite crystal, whereas the linear species are just small enough to enter. The ratio of intracrystalline diffusivities is therefore effectively infinite, and a very clean separation is possible.

III. MICROPOROUS ADSORBENTS

Since adsorption is essentially a surface phenomenon, a practical adsorbent must have a high specific surface area, which means small diameter pores. Conventional adsorbents such as porous alumina, silica gel, and activated carbon have relatively wide pore size distributions, spanning the entire range from a few angstroms to perhaps 1 μm . For convenience the pores are sometimes divided into three classes:

Micropores:	<20 Å diameter
Mesopores:	20–500 Å diameter
Macropores:	>500 Å diameter

A diameter of 20 Å represents approximately the limiting pore size that can be measured by mercury intrusion. In pores smaller than this, transport becomes increasingly affected by molecule–pore wall interactions, and conventional theories based on molecular and Knudsen diffusion break down. The classification is somewhat arbitrary, however, since the point at which such effects become important also depends on the size of the diffusing molecule. Adsorption equilibrium in microporous adsorbents also depends to some extent on the pore size as well as on the nature of the surface, so control of the pore size distribution is important in the manufacture of an adsorbent for a particular separation.

Activated carbon is by far the most widely used adsorbent. It is available in a wide range of different forms that differ mainly in pore size and pore size distribution. The carbon surface is essentially nonpolar although some polarity can be imparted by surface oxidation or other pre-treatments. It is widely used for removal of low concentrations of organics, either from aqueous streams (for example, decolorization of sugar or water treatment) or from vapor streams (for example, in range hoods and other pollution-control devices). Crystalline silica adsorbents such as silicalite are also organophilic but are substantially more expensive than activated carbon so their application is generally limited to situations where, for some reason, the use of carbon is not appropriate.

In “molecular sieve” adsorbents, such as zeolites and carbon molecular sieves, the micropore size distribution is extremely narrow, thus allowing the possibility of kinetic separations based on differences in molecular size. However, this feature is utilized in only a few commercial adsorption separation processes, and in the majority of such processes the separation depends on differences in the adsorption equilibrium rather than on the kinetics, even though a “molecular sieve” adsorbent may be used.

The Al-rich (cationic) zeolites have highly polar internal surfaces. The polarity increases with increasing cation charge and decreasing cation size. However, the relationship between the nature of the cation and the surface properties is complex because the differences in cation location (sites) must also be considered.

The commercially available zeolite adsorbents consist of small microporous zeolite crystals, aggregated with the aid of a clay binder. The pore size distribution thus has a well-defined bimodal character, with the diameter of the intracrystalline micropores being determined by the crystal structure and the macropore size being determined by the crystal diameter and the method of pelletization. As originally defined, the term *zeolite* was restricted to aluminosilicate structures, which can be regarded as assemblages of SiO_2 and AlO_2 tetrahedra. However, essentially

TABLE I Some Important Applications of Zeolite Adsorbents

Framework	Cationic form	Formula of typical unit cell	Window	Effective channel diameter (Å)	Application
A	Na	Na ₁₂ [(AlO ₂) ₁₂ (SiO ₂) ₁₂]	8-Ring (obstructed)	3.8	Desiccant: CO ₂ removal from natural gas
	Ca	Ca ₅ Na ₂ [(AlO ₂) ₁₂ (SiO ₂) ₁₂]	8-Ring (free)	4.4	Linear paraffin separation; air separation
	K	K ₁₂ [(AlO ₂) ₁₂ (SiO ₂) ₁₂]	8-Ring (obstructed)	2.9	Drying of cracked gas containing C ₂ H ₄ , etc.
X	Li(LSX)	Li ₉₆ [(AlO ₂) ₉₆ (SiO ₂) ₉₆]	12-Ring	8.4	PSA oxygen production
	Na	Na ₈₆ [(AlO ₂) ₈₆ (SiO ₂) ₁₀₆]	12-Ring	8.4	Pressure swing H ₂ purification
	Ca	Ca ₄₀ Na ₆ [(AlO ₂) ₈₆ (SiO ₂) ₁₀₆]	12-Ring	8.0	Removal of mercaptans from natural gas
	Sr, Ba ^a	Sr ₂₁ Ba ₂₂ [(AlO ₂) ₈₆ (SiO ₂) ₁₀₆]	12-Ring	8.0	Xylene separation
Y	K	K ₅₆ [(AlO ₂) ₅₆ (SiO ₂) ₁₃₆]	12-Ring	8.0	Xylene separation
	Ca	Ca ₂₈ [(AlO ₂) ₅₆ (SiO ₂) ₁₃₆]	12-Ring	8.0	Fructose–glucose separation
Mordenite	Ag	Ag ₈ [(AlO ₂) ₈ (SiO ₂) ₄₀]	12-Ring	7.0	I ₂ and Kr removal from nuclear off-gases
	H	H ₈ [(AlO ₂) ₈ (SiO ₂) ₄₀]			
Silicalite	—	(SiO ₂) ₉₆	10-Ring	6.0	Removal of organic compounds from water
ZSM-5	Na	Na ₃ [(AlO ₂) ₃ (SiO ₂) ₉₃]	10-Ring	6.0	Xylene separation

^a Also K–BaX.

pure silica analogs of many zeolite structures, as well as topologically similar AlPO₄ structures (AlPO₄ sieves), have now been prepared, and for practical purposes it is therefore convenient to consider such materials zeolites even though they do not fall within the traditional definition of a zeolite. Examples of some practically important zeolite adsorbents are given in Table I, together with the nominal micropore diameters, as determined from the crystal structures.

Carbon molecular sieves are produced by controlled pyrolysis and subsequent oxidation of coal, anthracite, or organic polymer materials. They differ from zeolites in that the micropores are not determined by the crystal structure and there is therefore always some distribution of micropore size. However, by careful control of the manufacturing process the micropore size distribution can be kept surprisingly narrow, so that efficient size-selective adsorption separations are possible with such adsorbents. Carbon molecular sieves also have a well-defined bimodal (macropore–micropore) size distribution, so there are many similarities between the adsorption kinetic behavior of zeolitic and carbon molecular sieve systems.

IV. ADSORPTION EQUILIBRIUM

A. Thermodynamics of Adsorption

At sufficiently low concentrations on a homogeneous surface the equilibrium isotherm for physical adsorption will always approach linearity (Henry's law). The limiting slope of the isotherm [$\lim_{p \rightarrow 0}(\partial q / \partial p)_T$] is referred to as the Henry constant K' . It is evident that the Henry con-

stant is simply a thermodynamic equilibrium constant, and the temperature dependence therefore follows the familiar van Hoff equation,

$$K' = K'_0 e^{-\Delta H_0 / RT} \quad (1)$$

where $-\Delta H_0$ is the limiting heat of adsorption at low coverage, R the gas constant, and T absolute temperature. Since adsorption is generally exothermic, the Henry constant decreases with temperature. A corresponding dimensionless Henry constant K can be defined in terms of fluid-phase concentration c [$K = \lim_{c \rightarrow 0}(\partial q / \partial c)_T$], where q is the sorbate concentration in adsorbed phase, rather than partial pressure, and since for an ideal vapor phase $c = p/RT$, the two constants are related by $K = RTK'$. Henry's law corresponds physically to the situation where the adsorbed layer is so dilute that there is neither competition for adsorption sites nor sorbate–sorbate interaction. At higher concentration levels both of these effects become important.

The equilibrium isotherms for microporous adsorbents are generally of type I form in Brunauer's classification (Fig. 1). Such isotherms are commonly represented by the Langmuir model,

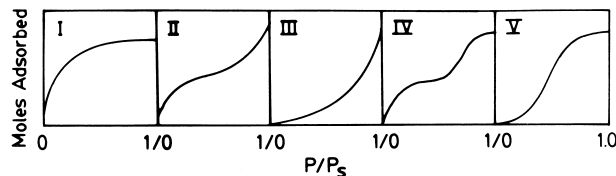


FIGURE 1 Brunauer's classification of equilibrium isotherms. P , sorbate pressure; P_s , saturation vapor pressure.

$$q/q_s = bp/(1 + bp) \quad (2)$$

where q_s is the saturation capacity and b an equilibrium constant that is directly related to the Henry constant ($K' = bq_s$). To a first approximation q_s is independent of temperature, so the temperature dependence of b is the same as that of the Henry constant [Eq. (1)]. The Langmuir model was originally derived for localized chemisorption on an ideal surface with no interaction between adsorbed molecules, but with certain approximations the same form of equation can be derived for mobile physical adsorption at moderate coverage. Although this model provides a quantitatively accurate description of the isotherms for only a few systems, the expression shows the correct asymptotic behavior at both high and low concentrations and therefore provides a useful qualitative or semiquantitative representation for many systems. A variety of more sophisticated model isotherms have been developed to take account of such factors as energetic heterogeneity and sorbate-sorbate interactions, but none of these has proved universally applicable. From the perspective of the overall modeling and design of adsorption systems, the more sophisticated models offer little advantage over the simple Langmuir model since any increase in accuracy is generally more than offset by the additional complexity of the model and the need for more empirical parameters.

When the equilibrium constant b is large (highly favorable adsorption) the Langmuir isotherm approaches irreversible or rectangular form,

$$p = 0, \quad q^* = 0; \quad p > 0, \quad q^* = q_s \quad (3)$$

where q^* represents the equilibrium constant ratio in the adsorbed phase. This provides the basis for a very useful limiting case, which is widely used in the analysis of adsorption column dynamics since the solutions for a rectangular isotherm are generally relatively simple and they provide a reasonably reliable prediction of the behaviour that can be expected for a real system when the isotherm is highly favorable.

According to the Langmuir model the heat of adsorption should be independent of adsorbed-phase concentration, but in practice the heat of adsorption generally varies quite significantly. For nonpolar sorbates an increase in the heat of sorption with coverage is generally observed, and this is commonly attributed to sorbate-sorbate interaction. For polar sorbates on polar adsorbents, the heat of sorption generally decreases with coverage, reflecting the dominance of energetic heterogeneity and the decreasing contribution of electrostatic contributions to the energy of adsorption at higher coverage (Fig. 2).

In homologous series such as the n -paraffins heats of adsorption increase regularly with carbon number (Fig. 3).

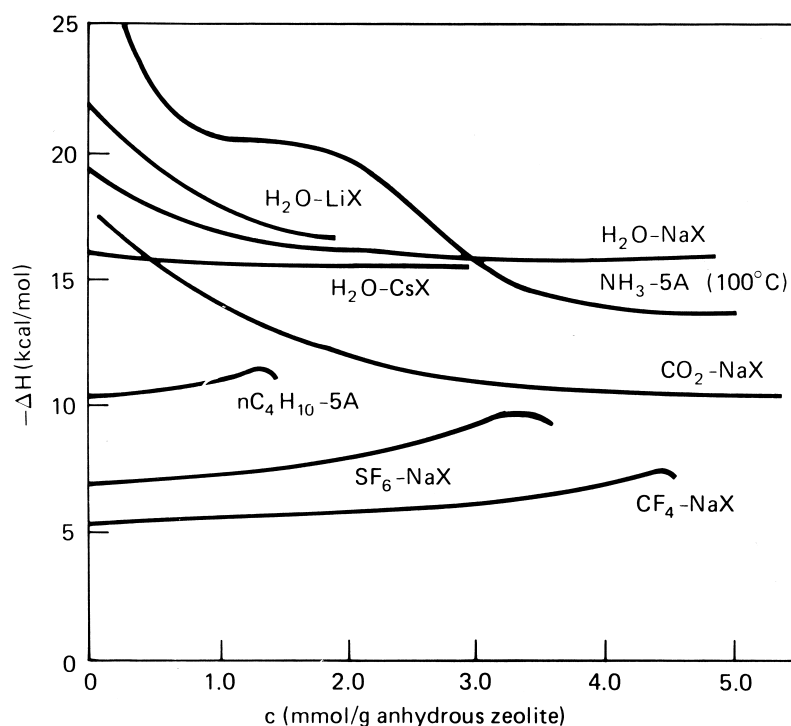


FIGURE 2 Variation of isosteric heat of sorption $-\Delta H_0$ with coverage c showing the difference in trends between polar and nonpolar sorbates. (Reprinted from Ruthven, D. M. (1976). *Sep. Purif. Methods* 5 (2), 184, copyright Marcel Dekker, Inc., New York.)

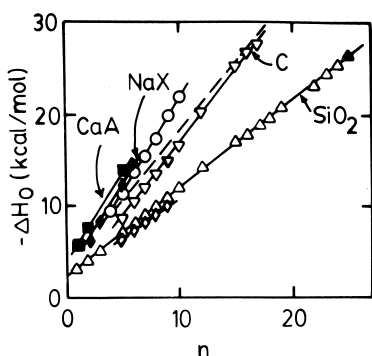


FIGURE 3 Variation of limiting heat of sorption ($-\Delta H_0$) with chain length n for homologous series of linear paraffins.

For more complex molecules a reasonable estimate of the heat of sorption can sometimes be made by considering “group contributions.” Such an approach works best for nonpolar sorbates on nonpolar surfaces but is subject to considerable error for polar systems in which the electrostatic energies of adsorption are large.

B. Adsorption of Mixtures

The Langmuir equation can be easily extended to multicomponent adsorption, for example, for a binary mixture of components 1 and 2:

$$\frac{q_1}{q_{s1}} = \frac{b_1 c_1}{1 + b_1 c_1 + b_2 c_2} \quad (4)$$

$$\frac{q_2}{q_{s2}} = \frac{b_2 c_2}{1 + b_1 c_1 + b_2 c_2}$$

Thermodynamic consistency requires that q_{s1} be equal to q_{s2} , but it is common practice to ignore this requirement, thereby introducing an additional parameter. This is legitimate if the equations are to be used purely as an empirical correlation, but it should be recognized that since thermodynamic consistency is violated such expressions are not valid over the entire composition range.

For an equilibrium-based separation process a convenient measure of the intrinsic selectivity of the adsorbent is the separation factor α_{12} , defined by analogy with relative volatility as:

$$\alpha_{12} = (X_1/Y_1)/(X_2/Y_2) \quad (5)$$

where X and Y are the mole fraction in the adsorbed and fluid phases, respectively. For a system that obeys the binary Langmuir isotherm [Eq. (4)] it is evident that α_{12} ($=b_1/b_2$) is independent of concentration. An approximate estimate of the separation factor can therefore be derived from the ratio of the Henry's law constants. A constant separation factor simplifies considerably the problem of modeling the adsorption process, and the Langmuir

model is therefore very useful for developing an initial understanding of the system dynamics and for preliminary design. However, the inherent limitations of such a model should be clearly recognized.

Although the multicomponent Langmuir equations account qualitatively for competitive adsorption of the mixture components, few real systems conform quantitatively to this simple model. For example, in real systems the separation factor is generally concentration dependent, and azeotrope formation ($\alpha = 1.0$) and selectivity reversal (α varying from less than 1.0 to more than 1.0 over the composition range) are relatively common. Such behavior may limit the product purity attainable in a particular adsorption separation. It is sometimes possible to avoid such problems by introducing an additional component into the system which will modify the equilibrium behavior and eliminate the selectivity reversal.

The problem of predicting multicomponent adsorption equilibria from single-component isotherm data has attracted considerable attention, and several more sophisticated approaches have been developed, including the ideal adsorbed solution theory and the vacancy solution theory. These theories provide useful quantitative correlations for a number of binary and ternary systems, although available experimental data are somewhat limited. A simpler but purely empirical approach is to use a modified form of isotherm expression based on Langmuir–Freundlich or “loading ratio correlation” equations:

$$\frac{q_1}{q_s} = \frac{b_1 p_1^{n_1}}{1 + b_1 p_1^{n_1} + b_2 p_2^{n_2}} \quad (6)$$

$$\frac{q_2}{q_s} = \frac{b_2 p_2^{n_2}}{1 + b_1 p_1^{n_1} + b_2 p_2^{n_2}}$$

From the perspective of the design engineer, the advantage of this approach is that the expressions for the adsorbed-phase concentrations are simple and explicit. However, the expressions do not reduce to Henry's law in the low-concentration limit, which is a thermodynamic requirement for physical adsorption. They therefore suffer from the disadvantage of any purely empirical equations, and they do not provide a reliable basis for extrapolation outside the range of experimental study.

V. ADSORPTION KINETICS

Physical adsorption at a surface is extremely rapid, and the kinetics of physical adsorption are invariably controlled by mass or heat transfer rather than by the intrinsic rate of the surface process. Biporous adsorbents such as pelleted zeolites or carbon molecular sieves offer three distinct resistances to mass transfer: the external resistance of the

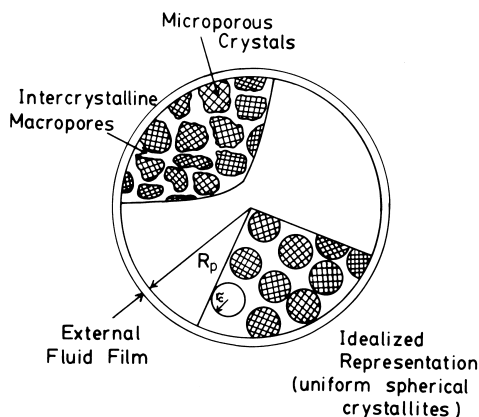


FIGURE 4 Schematic diagram of a biporous adsorbent pellet showing the three resistances to mass transfer (external fluid film, macropore diffusion, and micropore diffusion). R_p pellet radius; r_c crystal radius.

fluid film, the diffusional resistance associated with transport through the macropores, and the intracrystalline or micropore diffusional resistance (Fig. 4). Depending on the particular system and the conditions, any one of these resistances may be rate controlling, or the rate may be determined by the combined effects of more than one mass transfer resistance.

A. Micropore Diffusion

It is convenient to correlate transport data in terms of a diffusivity defined according to Fick's first equation,

$$J = -D(c) \frac{\partial c}{\partial z} \quad (7)$$

where J is flux, D diffusivity, c fluid-phase concentration, and z the distance. The true driving force for any transport process is, however, the gradient of the chemical potential rather than the concentration gradient, so one can write, more generally,

$$J = -Bc \frac{\partial \mu}{\partial z} \quad (8)$$

where B is mobility and μ the chemical potential. By considering equilibrium with an ideal vapor phase, it may be shown that the Fickian diffusivity (D) and the thermodynamic mobility (B) are related by:

$$D = BRT \frac{d \ln a}{d \ln c} = D_0 \frac{d \ln p}{d \ln c} \quad (9)$$

where c is the absorbed phase concentration, p the partial pressure, and the limiting diffusivity $D_0 = BRT$. It is evident that for an ideal system (activity proportional to concentration) Eq. (9) reduces to the Fickian formulation with $D = D_0$. However, for a nonideal system the factor

$d \ln p / d \ln c$ may be very different from unity. Such considerations apply equally to diffusion in liquids or gases as well as to diffusion in an adsorbed phase. However, for gaseous systems the deviations from ideality are generally small, and even for liquid systems the deviations from Henry's law are often modest over substantial ranges of concentration. By contrast, for an adsorbed phase the relationship between activity and concentration (the equilibrium isotherm) is almost always highly nonlinear.

The factor $d \ln p / d \ln q$ approaches unity in the Henry's law region and infinity in the saturation region of the isotherm, so a strong concentration dependence of the Fickian diffusivity (D increasing with q) is to be expected. For example, for a Langmuir system,

$$\frac{d \ln p}{d \ln q} = \frac{1}{1 - q/q_s}; \quad D = \frac{D_0}{1 - q/q_s} \quad (10)$$

In principle the mobility B and therefore the corrected diffusivity D_0 are also concentration-dependent, so Eq. (12) does not necessarily predict quantitatively the concentration dependence of D even for a system where the isotherm obeys the Langmuir equation. Nevertheless, the concentration dependence of B is generally modest compared with that of the thermodynamic factor, so a monotonic increase in diffusivity with adsorbed-phase concentration is commonly observed (Fig. 5). Clearly in any attempt to relate transport properties to the physical properties of the system it is important to examine the corrected, diffusivity D_0 (or the mobility B) rather than the Fickian diffusivity, which is in fact a product of kinetic and thermodynamic factors.

Micropore diffusion differs in several important respects from diffusion in macropores or in bulk fluids since the diffusing molecule never escapes from the force field of the solid. Under these conditions repulsive interactions are important, and relatively large differences in diffusivity may therefore occur between different stereoisomers, reflecting differences in molecular shape. Furthermore, small changes in pore diameter can affect the diffusivity by orders of magnitude, and on this basis a suitable adsorbent may sometimes be tailored to provide a high kinetic selectivity between similar molecules. The most important practical example is the separation of oxygen and nitrogen on a carbon molecular sieve.

Micropore diffusion is an activated process, and the temperature dependence can generally be correlated according to an Eyring equation,

$$D_0 = D_* e^{-E/RT} \quad (11)$$

where D_* is a pre-exponential factor and E the diffusional activation energy. The diffusional activation energy is a useful property which for a given sorbate-sorbent system

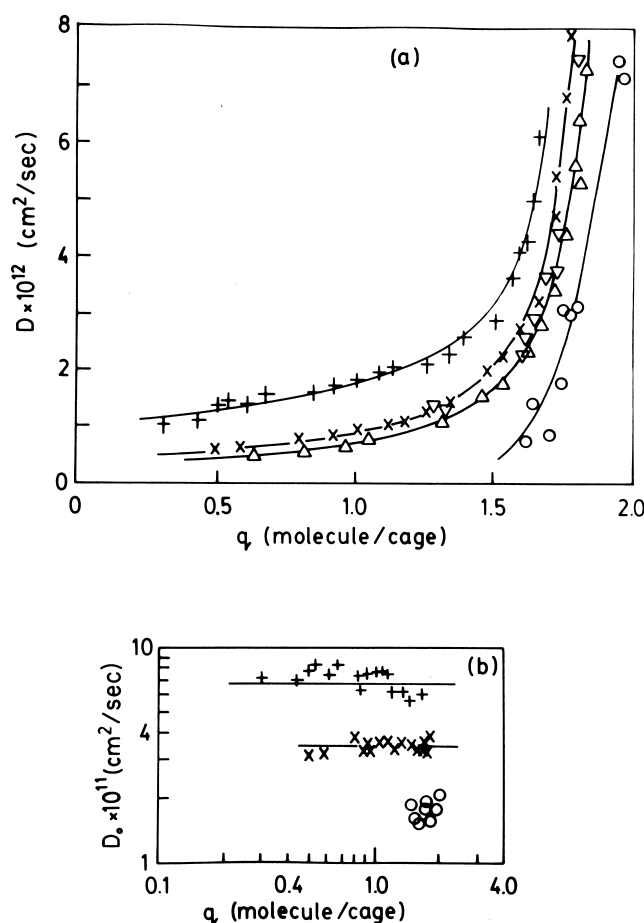


FIGURE 5 Variation of (a) intracrystalline diffusivity and (b) corrected diffusivity D_0 [Eq. (12)] with sorbate concentration q for n -heptane in Linde 5A zeolite crystals. $\circ$, 409 K; Δ , ∇ 439 K (adsorbent, desorbent, respectively); $\times$, 462 K; $+$, 491 K. (Reproduced by permission of the National Research Council of Canada from Ruthven, D. M., and Doetsch, I. H. (1974). *Can. J. Chem.* **52**, 2722.)

is commonly more constant than the actual value of the diffusivity. For zeolite adsorbents the variation of diffusional activation energy with molecular size and shape has been examined in considerable detail.

Many practical adsorption processes involve multicomponent systems, so the problem of micropore diffusion in a mixed adsorbed phase is both practically and theoretically important. Major progress in understanding the interaction effects has been achieved by Krishna and his coworkers through the application of the Stefan-Maxwell approach. The diverse patterns of concentration dependence of diffusivity that have been observed for many systems can, in most cases, be understood on this basis. The reader is referred, for details, to the review articles cited in the bibliography.

B. Macropore Diffusion

Diffusion in macropores occurs mainly by the combined effects of bulk molecular diffusion (as in the free fluid) and Knudsen flow, with generally smaller contributions from other mechanisms such as surface diffusion and Poiseuille flow. Knudsen flow, which has the characteristics of a diffusive process, occurs because molecules striking the pore wall are instantaneously adsorbed and re-emitted in a random direction. The relative importance of bulk and Knudsen diffusion depends on the relative frequency of molecule-molecule and molecule-wall collisions, which in turn depends on the ratio of the mean free path to pore diameter. Thus Knudsen flow becomes dominant in small pores at low pressures, while in larger pores and at higher pressures diffusion occurs mainly by the molecular mechanism. Since the mechanism of diffusion may well be different at different pressures, one must be cautious about extrapolating from experimental diffusivity data, obtained at low pressures, to the high pressures commonly employed in industrial processes.

The combined effects of Knudsen, D_K , and molecular (fluid-phase) diffusion D_m are commonly estimated from the expression:

$$\frac{1}{D_p} = \tau \left(\frac{1}{D_m} + \frac{1}{D_K} \right) \quad (12)$$

where τ is an empirical factor, characteristic of the adsorbent, that corrects for the effects of pore direction and nonuniform pore diameter. Modeling the pore structure as a three-dimensional assemblage of uniform, randomly oriented cylinders suggests a value of $\tau = 3$, and experimental values are typically within the range 2–4.

Since the transport processes within macropores are fairly well understood, it is generally possible to make a reasonable *a priori* estimate of the effective macropore diffusivity, at least within a factor of ~ 2 .

C. External Mass Transfer Resistance

External mass transfer rates are generally correlated in terms of a linear driving force expressions,

$$\partial q / \partial t = k_f a (c - c^*) \quad (13)$$

where t is time, k_f the external mass coefficient, and c^* the equilibrium value of c . Mass transfer rates in packed beds have been measured extensively, and the subject has generated considerable controversy in the literature. However, the matter has now been settled due largely to the diligent work of Wakao and collaborators. It appears that in many of the earlier measurements the effects of axial mixing were underestimated, leading to erroneously low apparent values for the film coefficient k_f . By taking proper account

of axial dispersion Wakao was able to correlate many of the data from different laboratories for both gas and liquid systems in accordance with the following correlation for the Sherwood number:

$$Sh = \frac{2k_f R_p}{D_m} = 2.0 + 1.1Sc^{1/3}Re^{1/2} \quad (14)$$

where Sc is the Schmidt number and Re the Reynolds number (based on particle diameter). However, it should be recognized that if this correlation is used to estimate the film coefficient it is essential also to use a realistic value for the axial dispersion coefficient. Otherwise, the combined effects of external mass transfer resistance and axial mixing will be underestimated.

D. Overall Mass Transfer Resistance

It has been well established that the kinetics of a diffusion-controlled process can be approximately represented by a linearized rate expression of the form:

$$\partial \bar{q} / \partial t = k(q^* - \bar{q}) \quad (15)$$

where the effective rate constant k is related to the diffusional time constant by $k \approx 15D/r^2$ (r being the particle radius), and $\bar{q}$ is the value of q averaged over a particle. This approximation, due originally to Glueck, is at its best for linear equilibrium systems and long adsorption columns, and it is at its worst when the isotherm is rectangular and for very short columns or single particles. When several resistances to mass transfer are significant (as in Fig. 4), the overall rate constant is given approximately by the reciprocal addition rule:

$$\frac{1}{kK} = \frac{R_p}{3k_f} + \frac{R_p^2}{15\varepsilon_p D_p} + \frac{r_c^2}{15KD_c} \quad (16)$$

where ε_p is the macroporosity of the adsorbent particle and D_c the intracrystalline (micropore) diffusivity. These approximations are especially useful in the modeling of adsorption column dynamics for more complex nonisothermal and multicomponent systems, since the replacement of a diffusion equation by a simple linearized rate expression leads to a general reduction in mathematical complexity and a corresponding reduction in the computer time requirement. The rigorous solution of diffusion equation models is generally not practically feasible except for the simplest systems.

E. Measurement of Intraparticle Diffusivities

The customary way of measuring intraparticle macropore diffusivities is the Wicke–Kallenbach method, which depends on measuring the flux through a pellet under steady-state conditions when the two faces are maintained at

known concentrations. The same method has also been adapted to the measurement of micropore diffusion in large crystals of certain zeolites.

Alternatively one can in principle derive both micropore and macropore diffusivities from measurements of the transient uptake rate for a particle (or assemblage of crystals) subjected to a step change in ambient sorbate pressure or concentration. The main problem with this approach is that the overall uptake rate may be controlled by several different processes, including both heat and extraparticle mass transfer as well as intraparticle or intracrystalline diffusion. The intrusion of such rate processes is not always obvious from a cursory examination of the experimental data, and the literature of the subject is replete with incorrect diffusivities (usually erroneously low values) obtained as a result of intrusion of such extraneous effects. Nevertheless, provided that intraparticle diffusion is sufficiently slow, the method offers a useful practical alternative to the Wicke–Kallenbach method.

Chromatographic methods offer a useful alternative to conventional batch uptake rate measurements. The advantage of these methods is that heat transfer effects can be greatly reduced and in most cases eliminated by the use of a high carrier flow rate and a low sorbate concentration. The main disadvantage is that the broadening of the response peak results from the combined effects of axial dispersion and mass transfer resistance. It is therefore necessary either to eliminate or to allow for axial dispersion in the column, and this is often more difficult than it may at first sight appear. Nevertheless, the method is quick and straightforward and requires no special equipment. It is therefore especially useful for preliminary adsorbent-screening studies when a rapid means of obtaining approximate kinetic and equilibrium data is required.

In the zero length column (ZLC) method, which can be regarded as a derivative of the traditional chromatographic method, a small sample of adsorbent is pre-equilibrated with the sorbate under well-defined conditions and then purged, at a constant flow rate, with an inert (nonadsorbing) gas (usually He), monitoring continuously the composition of the effluent stream. From analysis of the ZLC desorption curve both the adsorption equilibrium constant and the internal diffusivity can be obtained. The method retains the advantages of the traditional chromatographic method while eliminating the need to account for axial dispersion.

A more sophisticated method which has found wide application in the study of intracrystalline diffusion in zeolites is the nuclear magnetic resonance (NMR) pulsed field gradient self-diffusion method. The method, which is limited to hydrocarbons and other sorbates with a sufficient density of unpaired nuclear spins, depends on measuring directly the mean square distance traveled by molecules,

tagged according to the phase of their nuclear spins, during a known time interval of a few milliseconds. The quantity measured is thus the self-diffusivity D_s rather than the transport diffusivity, since under the conditions of the experiment there is no concentration gradient. The two quantities are related, however, by a well-defined relationship, which in the Henry's law region reduces simply to $D_s = D_0 = \lim_{c \rightarrow 0} D$.

VI. ADSORPTION COLUMN DYNAMICS

In most adsorption processes the adsorbent is contacted with fluid in a packed bed. The analysis and rational design of such processes therefore require an understanding of the dynamic behavior of such systems. What is required is a mathematical model which will allow the effluent concentration to be predicted for any defined change in feed concentration, but two simple situations are of special interest:

1. The response of a column, initially at equilibrium with the feed stream, to a step change in the concentration of an adsorbable species in the feed. This is referred to as the breakthrough curve (for a concentration increase) or the desorption curve (for a concentration decrease). The simplest case is a clean bed exposed to a finite steady feed concentration at time zero (or the corresponding desorption step), but changes between two finite concentrations can also be considered in the same way. The breakthrough curve clearly gives directly the breakthrough time (i.e., the time at which the effluent concentration reaches the maximum allowable level in a purification process) and hence the dynamic capacity of the bed.
2. The response of a column to a pulse injection of sorbate into an inert (nonadsorbing) carrier. This is referred to as the chromatographic response. Such measurements provide a convenient way of determining kinetic and equilibrium data.

For a linear system essentially the same information can be deduced from either a pulse or step response measurement. (Since the pulse is the time derivative of the step function, the response to the pulse will be the derivative of the step response.) Both methods are widely used, and the choice is therefore dictated by experimental convenience rather than by fundamental theoretical considerations.

The broad features of the dynamic response are determined by the form of the equilibrium isotherm. The behavior may be significantly modified by kinetic effects, but the general pattern of the system response remains the same even when resistance to mass transfer is impor-

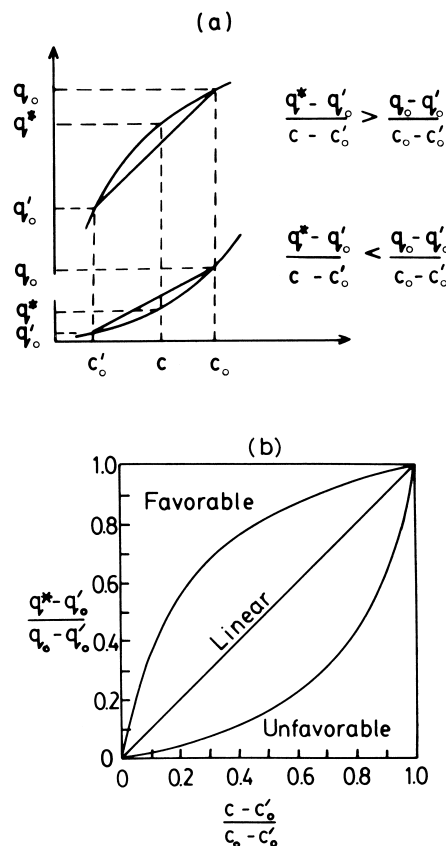


FIGURE 6 (a) Equilibrium isotherms and (b) dimensionless equilibrium diagram showing distinction between favorable, unfavorable, and linear systems. (Reprinted with permission from Ruthven, D. M. (1984). "Principles of Adsorption and Adsorption Processes," copyright John Wiley & Sons, New York.)

tant. This means that a useful qualitative understanding can be achieved simply from equilibrium theory, and this approach has proved especially valuable for multicomponent systems where a more precise analysis including both kinetic and equilibrium effects is difficult.

Equilibrium isotherms can be classified as favorable or unfavorable according to the shape of the X - Y diagram (Fig. 6). It is evident that if an isotherm is favorable for adsorption, and that is the most common situation (corresponding to a type I isotherm of Brunauer's classification), it will be unfavorable for desorption. The rate at which a disturbance propagates through the column is determined by the slope of the equilibrium isotherm and, for a favorable isotherm, is higher at higher concentrations. This leads to "self-sharpening" of the concentration profile and, in a column of sufficient length, to "constant-pattern" behavior (Fig. 7). In the initial region of the column the concentration profile broadens as it progresses through the column, but after some distance a coherent dynamic situation is achieved in which the tendency for the

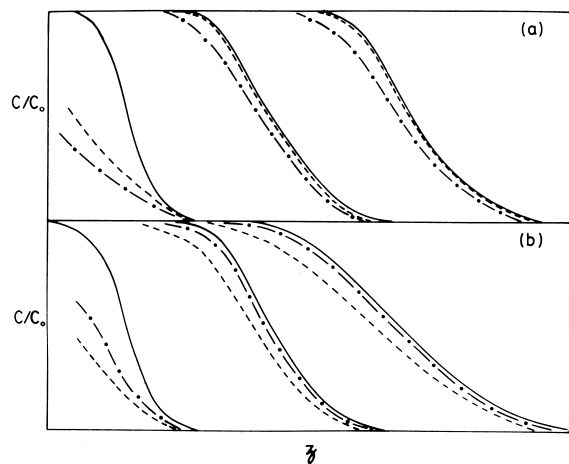


FIGURE 7 Schematic diagram showing (a) approach to constant-pattern behavior for a system with favorable equilibrium and (b) approach to proportionate-pattern limit for a system with unfavorable isotherm. Key: c/c_0 , —; q/q_0 , ---; c^*/c_0 , —·—. (Reprinted with permission from Ruthven, D. M. (1984). "Principles of Adsorption and Adsorption Processes," copyright John Wiley & Sons, New York.)

concentration front to broaden due to the effects of mass transfer resistance and axial dispersion is exactly balanced by the self-sharpening effect arising from the variation of the characteristic velocity and concentration. Once this state is reached the concentration profile propagates without further change in shape. This is the basis of the LUB (length of unused bed) method of adsorber design, which is considered in greater detail [see Eq. (25)].

In the case of an unfavorable isotherm (or equally for desorption with a favorable isotherm) a different type of behavior is observed. The concentration front or mass transfer zone, as it is sometimes called, broadens continuously as it progresses through the column, and in a sufficiently long column the spread of the profile becomes directly proportional to column length (proportionate pattern behavior). The difference between these two limiting types of behavior can be understood in terms of the relative positions of the gas, solid, and equilibrium profiles for favorable and unfavorable isotherms (Fig. 7).

A. Mathematical Modeling

The pattern of flow through a packed adsorbent bed can generally be described by the axial dispersed plug flow model. To predict the dynamic response of the column therefore requires the simultaneous solution, subject to the appropriate initial and boundary conditions, of the differential mass balance equations for an element of the column,

$$-D_L \frac{\partial^2 c_i}{\partial z^2} + \frac{\partial}{\partial z}(v c_i) + \frac{\partial c_i}{\partial t} + \left(\frac{1 - \varepsilon}{\varepsilon} \right) \frac{\partial \bar{q}_i}{\partial t} = 0 \quad (17)$$

(D_L is the axial dispersion coefficient, z distance, v the interstitial fluid velocity, and ε the voidage of the adsorbent bed) together with the adsorption rate expression for each component, which can be written in the general form:

$$\frac{\partial \bar{q}_i}{\partial t} = f(\bar{q}_i, q_s, \dots; c_i, c_j, \dots; T) \quad (18)$$

It should be understood that this rate expression may in fact represent a set of diffusion and mass transfer equations with their associated boundary conditions, rather than a simple explicit expression. In addition one may write a differential heat balance for a column element, which has the same general form as Eq. (17), and a heat balance for heat transfer between particle and fluid. In a nonisothermal system the heat and mass balance equations are therefore coupled through the temperature dependence of the rate of adsorption and the adsorption equilibrium, as expressed in Eq. (18).

Solving this set of equations is a difficult task, and some simplification is therefore generally needed. Some of the simplified systems for which more or less rigorous solutions have been obtained are summarized below.

For a system with n components (including nonadsorbable inert species) there are $n - 1$ differential mass balance equations of type (17) and $n - 1$ rate equations [Eq. (18)]. The solution to this set of equations is a set of $n - 1$ concentration fronts or mass transfer zones separated by plateau regions and with each mass transfer zone propagating through the column at its characteristic velocity as determined by the equilibrium relationship. In addition, if the system is nonisothermal, there will be the differential column heat balance and the particle heat balance equations, which are coupled to the adsorption rate equation through the temperature dependence of the rate and equilibrium constants. The solution for a nonisothermal system will therefore contain an additional mass transfer zone traveling with the characteristic velocity of the temperature front, which is determined by the heat capacities of adsorbent and fluid and the heat of adsorption. A nonisothermal or adiabatic system with n components will therefore have n transitions or mass transfer zones and as such can be considered formally similar to an $(n + 1)$ -component isothermal system.

The number of transitions or mass transfer zones provides a direct measure of the system complexity and therefore of the ease or difficulty with which the behavior can be modeled mathematically. It is therefore convenient to classify adsorption systems in the manner indicated in Section V.B. It is generally possible to develop full dynamic models only for the simpler classes of systems, involving one, two, or at the most three transitions.

B. Classification According to Number of Transitions

1. Single-Transition Systems

a. One adsorbable component plus inert carrier, isothermal operation.

i. Trace concentrations. If the concentration of adsorbable species is small, variation in flow rate through the column may be neglected. Equation (17) reduces to:

$$-D_L \frac{\partial^2 c}{\partial z^2} + \frac{v \partial c}{\partial z} + \frac{\partial c}{\partial t} + \left(\frac{1 - \varepsilon}{\varepsilon} \right) \frac{\partial \bar{q}}{\partial t} = 0 \quad (19)$$

If the equilibrium is linear, exact analytical solutions for the column response can be obtained even when the rate expression is quite complex. In most of the published solutions, axial dispersion is also neglected, but this simplification is not essential and a number of solutions including both axial dispersion and more than one diffusional resistance to mass transfer have been obtained. Analytical solutions can also be obtained for an irreversible isotherm with negligible axial dispersion, but the case of an irreversible isotherm with significant axial dispersion has not yet been solved analytically.

For nonlinear systems the solution of the governing equations must generally be obtained numerically, but such solutions can be obtained without undue difficulty for any desired rate expression with or without axial dispersion. The case of a Langmuir system with linear driving force rate expression and negligible axial dispersion is a special case that is amenable to analytical solution by an elegant nonlinear transformation.

ii. Nontrace concentration. If the concentration of the adsorbable species is large it is necessary to account for the variations in flow rate through the adsorbent bed. This introduces an additional equation, making the solution more difficult. Numerical solutions can still be obtained, but few if any analytical solutions have been found for such systems.

b. Two adsorbable components (no carrier), isothermal operation. This is a special case since, in the absence of a carrier, the rate equations for the two adsorbable species are coupled through the continuity equation so that a single mass transfer zone is still obtained. The case of tracer exchange is a particularly simple example of this type of system since the adsorption process then involves equimolar exchange and the solutions, even for a large concentration step, are formally the same as for a linear trace component system.

2. Two-Transition Systems

Such systems can be of any of the following types: (1) isothermal, two adsorbable components plus inert carrier;

- (2) isothermal, three adsorbable components, no carrier; (3) adiabatic, one adsorbable component plus inert carrier; (4) adiabatic, two adsorbable components, no carrier.

The only case for which analytical solutions have been obtained is (3) when the equilibrium isotherm is of rectangular form. For such systems the mass balance equation is not coupled to the heat balance, and the solution for the concentration profile is the same as for an isothermal system. There is thus only one concentration front, the second transition being a pure temperature transition with no change in concentration. Solutions for the other cases can be obtained numerically, provided that a simple linearized rate expression is used.

3. Multiple-Transition Systems

Only a few full dynamic solutions for systems with more than two transitions have been derived, and for multicomponent adiabatic systems equilibrium theory offers the only practical approach.

C. Chromatography

Measurement of the mean retention time and dispersion of a concentration perturbation passing through a packed adsorption column provides a useful method of determining kinetic and equilibrium parameters. The carrier should be inert, and the magnitude of the concentration change must be kept small to ensure linearity of the system.

The principle of the method may be illustrated by considering the response to the injection of a perfect pulse of sorbate at the column inlet at time zero. The mean retention time t is given by the first moment of the response peak and is related to the dimensionless Henry constant by:

$$\bar{t} \equiv \frac{\int_0^\infty ct \, dt}{\int_0^\infty c \, dt} = \frac{L}{v} \left[1 + \left(\frac{1 - \varepsilon}{\varepsilon} \right) K \right] \quad (20)$$

where L is column length. Dispersion of the response peak, which arises from the combined effects of axial dispersion and finite mass transfer resistance, is conveniently measured by the second moment σ^2 of the response:

$$\sigma^2 \equiv \frac{\int_0^\infty c(t - \bar{t})^2 dt}{\int_0^\infty c \, dt} \quad (21)$$

For a dispersed plug flow system with K large ($K \gg 1$) it can be shown that:

$$\frac{\sigma^2}{\bar{t}^2} = \frac{D_L}{vL} + \left(\frac{\varepsilon}{1 - \varepsilon} \right) \left(\frac{v}{L} \right) \frac{1}{kK} \quad (22)$$

where k is the overall mass transfer coefficient defined according to Eq. (15).

The relationship with the familiar van Deemter equation giving the HETP (height equivalent to a theoretical plate) as a function of gas velocity,

$$\text{HETP} \equiv \frac{\sigma^2 L}{\bar{t}^2} = \frac{A_1}{v} + A^2 + A_3 v \quad (23)$$

can be easily derived by substituting the approximate relationship $D_L \approx 0.7D_m + vR_p$ in Eq. (22), whence it follows that coefficient $A_1 \approx 1.4D_m$, $A_2 = 2R_p$, and $A_3 = 2\varepsilon/(1 - \varepsilon)kK$.

In the low-velocity region the axial dispersion coefficient D_L is approximately independent of gas velocity and Eq. (22) can be rearranged to give:

$$\frac{\sigma^2 L}{2\bar{t}^2 v} = \frac{D_L}{v^2} + \left(\frac{\varepsilon}{1 - \varepsilon} \right) \frac{1}{kK} \quad (24)$$

from which it is evident that a plot of $(\sigma^2 L/2\bar{t}^2 v)$ versus $1/v^2$ should be linear with slope D_L and intercept $\varepsilon/(1 - \varepsilon)kK$. This provides a simple means of separating the effects of axial dispersion and mass transfer resistance. The shape of the response peak is rather insensitive to the nature of the mass transfer resistance, however, so even by more sophisticated methods of analysis it is generally not possible to establish the relative importance of the individual mass transfer resistances except by varying the adsorbent particle size and/or crystal size.

VII. CYCLIC BATCH ADSORPTION PROCESSES

The general mode of operation of a cyclic batch adsorption process is illustrated in Fig. 8. In its simplest form such a process employs two adsorbent beds, each of which is alternately saturated and regenerated. During the saturation or adsorption cycle, adsorption is continued until the mass transfer zone has almost reached the bed outlet. At this point the beds are switched so that the spent bed is replaced by a freshly regenerated bed, while the more strongly absorbed species is removed from the spent bed in

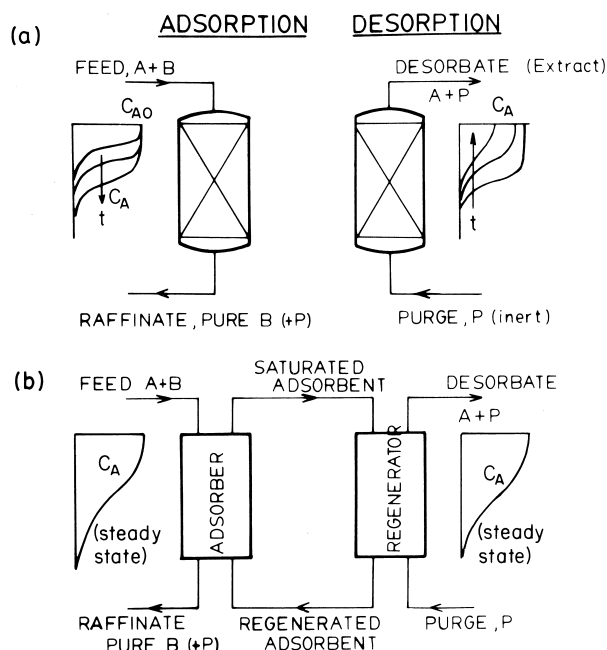


FIGURE 8 Schematic diagram showing the two basic modes of operating an adsorption separation process: (a) cyclic batch two-bed system; (b) continuous countercurrent system with adsorbent recirculation. Concentration profiles through the adsorbent bed are indicated. Component A is more strongly adsorbed than B. (Reprinted with permission from Ruthven, D. M. (1984). "Principles of Adsorption and Adsorption Processes," copyright John Wiley & Sons, New York.)

the regeneration desorption step. Some examples of such processes are given in Table II.

Processes of this type can be further classified according to the method used to regenerate the spent bed: thermal swing, pressure swing, purge gas stripping, or displacement desorption. In a thermal swing process desorption is accomplished by raising the temperature of the bed, either

TABLE II Examples of Cyclic Adsorption Separation Processes

Process	Liquid or gas phase ^a	Adsorbent	Selectivity	Regeneration method
Drying of gas streams	G	13X, 4A, or 3A molecular sieve	Equilibrium	Thermal swing or pressure swing
Drying of solvents	L	4A sieve	Equilibrium	Thermal swing
Solvent recovery	G	Activated carbon	Equilibrium	Steam stripping
H ₂ recovery	G	Molecular sieve	Equilibrium	Pressure swing
Air separation	G	Carbon molecular sieve	Kinetic	Pressure swing
		Zeolite	Equilibrium	
Linear paraffins separation	G	5A molecular sieve	Shape-selective sieving	Displacement or vacuum desorption
Wastewater purification	L	Activated carbon	Equilibrium	Steam stripping

^a Liquid; G, gas.

by heaters within the bed or, more commonly, by purging with a hot purge gas. At higher temperatures the adsorption equilibrium constant is reduced so that even quite strongly adsorbed species can be removed with a comparatively small purge gas volume. In a pressure swing process desorption is achieved simply by reducing the total pressure, while purge gas stripping depends on reducing the partial pressure by dilution with an inert purge gas. This generally requires a rather large purge volume, so such a process would normally be used only in special circumstances.

Displacement desorption is similar to purge gas stripping, except that an adsorbable species is used to displace the adsorbed component from the bed. The displacing component should be adsorbed somewhat less strongly than the preferentially adsorbed species so that the adsorption–desorption equilibrium can be shifted by varying the concentration of the desorbent. Such processes run more or less isothermally and offer a useful alternative to thermal swing processes for strongly adsorbed species when thermal swing would require temperatures high enough to cause cracking, coking, or rapid aging of the adsorbent. Steam stripping, which is widely used in solvent recovery systems, can be considered a combination of displacement desorption and thermal swing. The advantages and disadvantages of these methods of regeneration are summarized in Table III.

In general desorption is not carried to completion during the regeneration step, so the bed in fact operates between two partially loaded states. At the end of the desorption cycle the residue of the more strongly adsorbed species is concentrated near the bed outlet. If the same flow direction were maintained during adsorption this would cause contamination of the raffinate product at the beginning of the next adsorption step. This problem can be avoided by reversing the flow direction. An additional advantage

of reverse-flow regeneration is that the volume of purge required to regenerate the bed is reduced, so this mode of operation is almost always adopted.

Contact between the fluid phase and the solid adsorbent is generally accomplished in a packed adsorbent bed. A packed bed is simple and relatively inexpensive and it has good mass transfer characteristics. However, from the standpoint of pressure drop, and therefore power consumption, it is relatively inefficient. Such considerations become important when the throughput is large and the “value added” in the process is small. Examples include volatile organic compound (VOC) removal processes and desiccant cooling systems. For such systems a “parallel passage” contactor in which the adsorbent is in the form of a honeycomb, an array of parallel sheets, or a monolith, although more expensive in capital cost, proves to be a more economic option. Such adsorbents are commonly configured in the form of a slowly rotating wheel which allows the adsorbent to be exposed alternately to the feed streams and the regenerant or purge as it rotates. The regeneration section is often heated to yield the analog of a traditional thermal swing process.

A. Thermal Swing Processes

Cyclic thermal swing processes are widely used for purification operations such as drying or removal of CO₂ from natural gas. Design of a cyclic adsorption process requires knowledge of the dynamic capacity of the bed or the breakthrough curve. If mass transfer resistance and/or axial dispersion are significant, the dynamic capacity, which is determined by the extent to which the mass transfer front is broadened during passage through the column, may be much smaller than the static capacity determined from the equilibrium isotherm. If kinetic and equilibrium data are available and the system is sufficiently simple to

TABLE III Factors Governing Choice of Regeneration Method

Method	Advantages	Disadvantages
Thermal swing	Good for strongly adsorbed species, since small change in T gives large change in q^* ; desorbate can be recovered at high concentration; applicable to both gases and liquids	Thermal aging of adsorbent; heat loss means inefficiency in energy usage; unsuitable for rapid cycling, so adsorbent cannot be used with maximum efficiency; in liquid systems, high latent heat of interstitial liquid must be added
Pressure swing	Good where weakly adsorbed species is required in high purity; rapid cycling, efficient use of adsorbent	Very low pressure may be required; mechanical energy more expensive than heat; desorbate recovered at low purity
Displacement desorption	Good for strongly held species; avoids risk of cracking reactions during regeneration; avoids thermal aging of adsorbent	Product separation and recovery needed (choice of desorbent is crucial)

allow detailed mathematical modeling along the lines indicated in the previous sections, one can in principle predict the dynamic capacity for any defined feed and regeneration conditions. An *a priori* design of the bed is therefore feasible. Such an approach has been adopted only rather infrequently, however, probably because the capability of solving the governing equations for the more complex systems typical of industrial operations has been achieved only recently. A more common approach is to base the design on experimental measurements of dynamic capacity using the LUB concept. A breakthrough curve is measured using the same adsorbent under the same hydrodynamic conditions but in a laboratory-scale column. The LUB, which is essentially a measure of the width of the mass transfer zone, is given by

$$\text{LUB} = (1 - \bar{q}'/q_0)L = (1 - t'/\bar{t})L \quad (25)$$

where q_0 is the adsorbed-phase concentration in equilibrium with the feed, t' the break time, and $\bar{t}$ the mean intention time. These quantities can be calculated directly by integration from an experimental breakthrough curve (Fig. 9),

$$\bar{t} = \int_0^\infty (1 - c/c_0) dt$$

(striped area in Fig. 9)

$$t' = \int_0^{t'} (c - c/c_0) dt$$

(hatched area in Fig. 9)

where c_0 is the feed concentration of sorbate. The effective capacity of a column length L will be the equilibrium capacity of a column of length L' , where $L' = L - \text{LUB}$, and on this basis the size of a column required for a given duty can be readily estimated. It is important that the experimental LUB be measured under conditions that are precisely analogous to the large-scale process. For example, if the small laboratory column operates isothermally while the full-scale unit is adiabatic, the LUB may be seriously underestimated, leading to an inadequate design. Furthermore, the method is valid only for a constant-pattern system (adsorption with a favorable isotherm) and provides

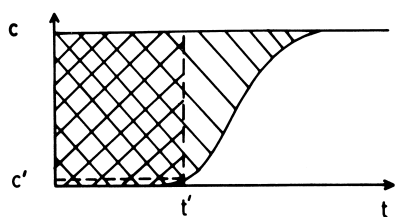


FIGURE 9 Sketch of a typical breakthrough curve showing relationship between break time t' and mean retention time $\bar{t}$.

no information on the regeneration conditions needed. In practice, in most two-bed purification processes the desorption step in fact controls the cycle, either directly or through the heat balance. Initial design of the regeneration cycle is commonly based on the assumption that during desorption the column approaches equilibrium. However, at the low concentrations prevailing during the later steps of desorption, kinetic effects may be important, so a more detailed analysis is desirable.

Another factor that is particularly important in the regeneration of molecular sieve driers is the rate at which the temperature is raised during regeneration. If this is too rapid relative to the rate of moisture removal, one may get rapid desorption of moisture from the initial section of the bed, which is in contact with the hot desorbent gas, followed by condensation of liquid water in the cooler regions some distance from the inlet, with serious consequences for adsorbent life.

To avoid the possibility of fluidizing the bed the system is normally operated in the downflow mode with upflow desorption since the gas velocity during desorption is normally lower than that during adsorption. The maximum upflow velocity is normally limited to 80% of the minimum fluidization velocity, while velocities as high as 1.8 times minimum fluidization can be tolerated in downflow.

B. Pressure Swing Processes

The general features of a simple two-bed pressure swing adsorption (PSA) system are shown in Fig. 10, and details of two simple cycles are shown in Fig. 11. One of the important features of such processes is that the less strongly adsorbed species (the raffinate product) can be recovered at high purity but at relatively low fractional recovery, while the more strongly adsorbed species (the extract product) is always recovered in impure form during the blowdown and purge steps. This type of process, is therefore especially suitable for gaseous separations when the feed is inexpensive and the less strongly adsorbed species is the required product. All three major industrial applications of PSA (air drying, air separation, and hydrogen purification) fulfill these requirements.

PSA systems are well suited to rapid cycling, making it possible to obtain relatively large throughput with relatively small adsorbent beds. However, the energy efficiency of such processes is not high, and since mechanical energy is generally more expensive than heat, PSA systems are generally not economic for large-scale operations. Their advantage lies in their compactness and simplicity, making them ideal for applications such as the production of medical oxygen in the home or in hospitals in remote areas. However, with recent improvements in process efficiency PSA processes are economically competitive with

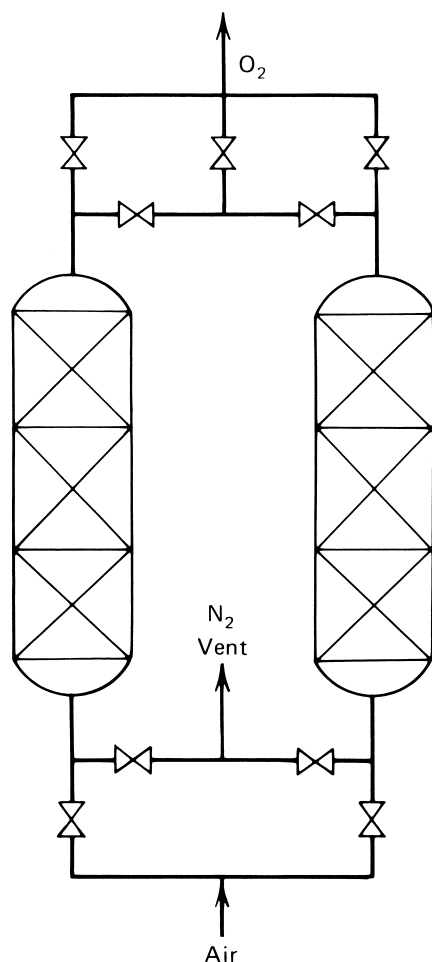


FIGURE 10 Schematic diagram of a simple two-bed pressure swing adsorption system.

cryogenic distillation for oxygen production rates up to about 250 tons/day.

Two types of PSA air-separation processes are in common use. When oxygen is the required product a nitrogen-selective zeolite adsorbent is used in order to produce oxygen as the (pure) raffinate product. Earlier processes generally used 5A or NaX zeolites operating between about 3 and 1 atm on a modified Skarstrom cycle (see Fig. 11a). However, most modern processes use LiX (highly exchanged low silica X), which has a much higher selectivity and capacity for nitrogen. The higher affinity for nitrogen makes it necessary to resort to vacuum desorption—sometimes called a vacuum swing cycle (VSA). A typical process operates with feed at about 1.2 atm and desorption at 0.3 atm. In large-scale units, a radial flow configuration is sometimes used in order to reduce pressure drop and thus reduce the power cost.

For nitrogen production, a carbon molecular sieve adsorbent is generally used. The equilibrium isotherms for oxygen and nitrogen on carbon molecular sieves are almost identical, but the micropore diffusivity of oxygen is much higher ($D_{O_2}/D_{N_2} \sim 30$). A kinetic separation is therefore possible, yielding nitrogen as the raffinate product. The process could be carried out in a Skarstrom cycle, but the cycle shown in Fig. 11(b) provides a more attractive alternative. This system is self-purging because the purge gas is provided by the residual nitrogen which desorbs during the “desorption” step. Although high-purity nitrogen can be obtained in this way, it is generally more economic to produce a nitrogen product of ~99% purity and remove the remaining oxygen by hydrogen addition and catalytic oxidation.

In the zeolite-based PSA process the argon is separated with the oxygen. For medical applications the presence of a small amount of argon is of little consequence, but it is a significant disadvantage for welding since the presence of even a small amount of argon leads to a significant reduction of flame temperature and cutting speed. In the carbon sieve process the argon and nitrogen are separated together as the raffinate product.

Although the simple two-bed PSA cycle is widely used in small-scale units, to achieve economic operations on a larger scale it is necessary to improve the energy efficiency of the process. This can be accomplished by using multiple-bed systems in which blowdown and repressurization take place in several stages in such a way that the high-pressure enriched gas at the end of the adsorption step in column 1 is used to pressurize partially column 2 and so on.

C. Displacement Desorption

One of the earliest and most successful processes for the separation of linear and branched-chain paraffins is the Exxon Ensorb process, shown schematically in Fig. 12. The process uses a 5A molecular sieve adsorbent, which admits the straight-chain paraffins but excludes the branched and cyclic isomers, with ammonia as the desorbent. The process operates isothermally at 550 to 600°F and essentially at atmospheric pressure with a cycle time that varies from about 12 to 30 min depending on the condition of the sieve and the linear-paraffin content of the feed. Other oil companies have similar processes. These differ mainly in the choice of desorbent, but ammonia is a particularly good choice since its high dipole moment allows it to compete with the much higher molecular weight paraffins while because of its low molecular weight and high volatility it is easily separated from the hydrocarbon products by flash distillation.

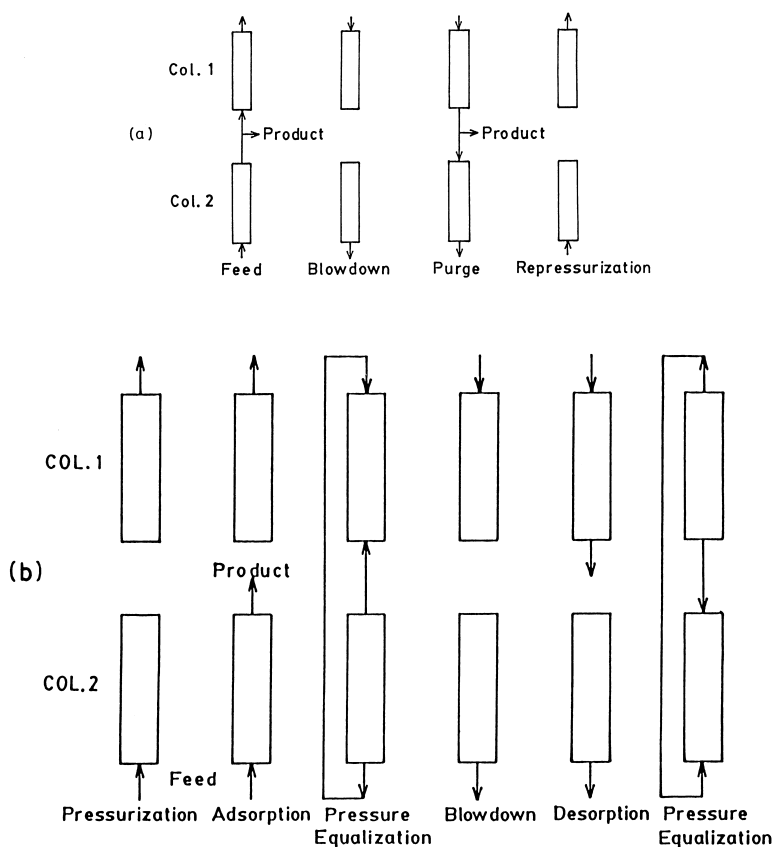


FIGURE 11 Sequence of steps in a two-bed pressure swing adsorption system. (a) Skarstrom cycle, (b) modified cycle for production of nitrogen using a carbon molecular sieve adsorbent.

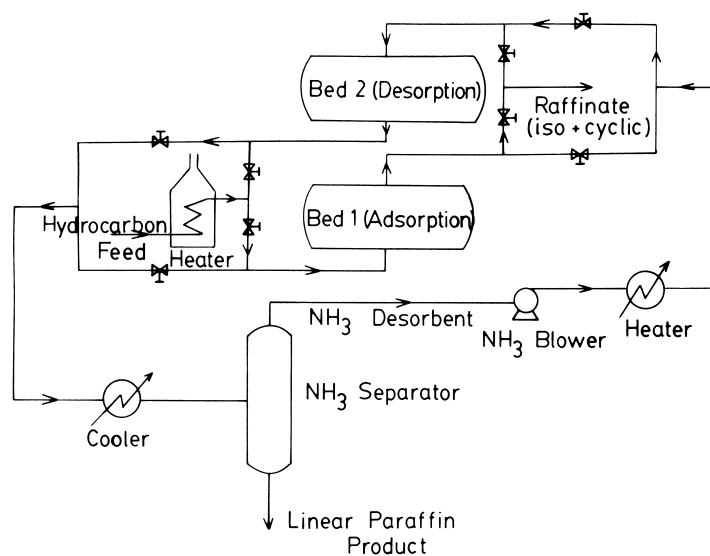


FIGURE 12 Schematic diagram of the Exxon Ensorb process. (Courtesy of Aromatics Technology Division of Exxon Chemical Company.)

VIII. CHROMATOGRAPHIC PROCESSES

It is well known to the analytical chemist that efficient separation of even rather similar compounds can be achieved in a chromatographic column. The possibility of scaling up such a process to preparative scale is inherently attractive, and many drugs, perfumes, and other compounds of high value are in fact separated in this way. However, such processes have generally been found unsuitable for the large-scale bulk separations typical of the petrochemical industry, and their practical usefulness is limited to systems with maximum throughputs of perhaps 1–2 tons/day. The main difficulty is that in large-diameter beds the HETP increases dramatically as a consequence of small nonuniformities in the packing, thus reducing the separation efficiency. Such effects can be minimized by very careful packing of the column but, even so, such processes are generally confined to high-value products and modest throughputs.

Production-scale chromatographs are generally operated under conditions somewhat different from those employed in analytical chromatography since the objective is to maximize throughput rather than resolution. As a result the column is generally operated at minimum resolution and under overload conditions. Feed pulses are injected successively so that the resolution between successive pulses is about the same as the resolution between the components of each pulse. Theoretical considerations suggest that for optimal design one should run six columns in parallel with feed switched in sequence to each column in such a way that the feed is injected into each column for one-sixth of the time with pure carrier flowing for five-sixths of the time.

IX. CONTINUOUS COUNTERCURRENT PROCESSES

The possibility of operating an adsorption separation as a continuous countercurrent process (Fig. 8), rather than in the cyclic batch mode, is theoretically attractive because countercurrent contact maximizes the driving force for mass transfer, thus providing more efficient utilization of the adsorbent than is possible in either cyclic batch or chromatographic systems. The main difficulty is that for countercurrent contact it is necessary either to circulate the adsorbent or, by appropriate design of the fluid flow system, to simulate adsorbent circulation. This makes the design of a countercurrent system more complex and reduces operational flexibility. For relatively easy separations (high separation factor, adequate mass transfer rates) the balance of economic advantage generally lies with a cyclic batch system, but for difficult separations in which selectivity is low or mass transfer slow the advantage of a continuous countercurrent system in reducing the required inventory of adsorbent must eventually outweigh the disadvantages of the more complex engineering.

A. Simulated Countercurrent Systems

Much of the benefit of countercurrent operation, without the problems associated with circulation of the adsorbent, can be achieved by using a multiple-column fixed-bed system with an appropriate sequence of column switching, designed to simulate a countercurrent flow system. The general scheme is illustrated in Fig. 13. Such systems are widely used in wastewater treatment,

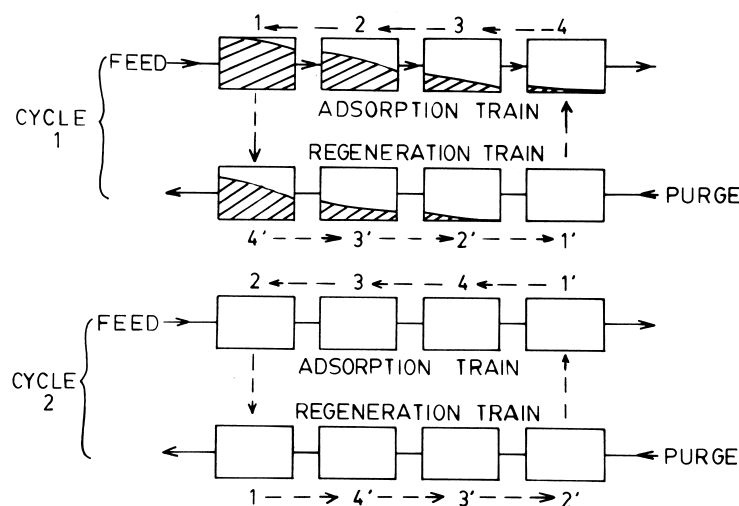


FIGURE 13 Schematic diagram showing the sequence of column interchange in a periodic countercurrent separation process.

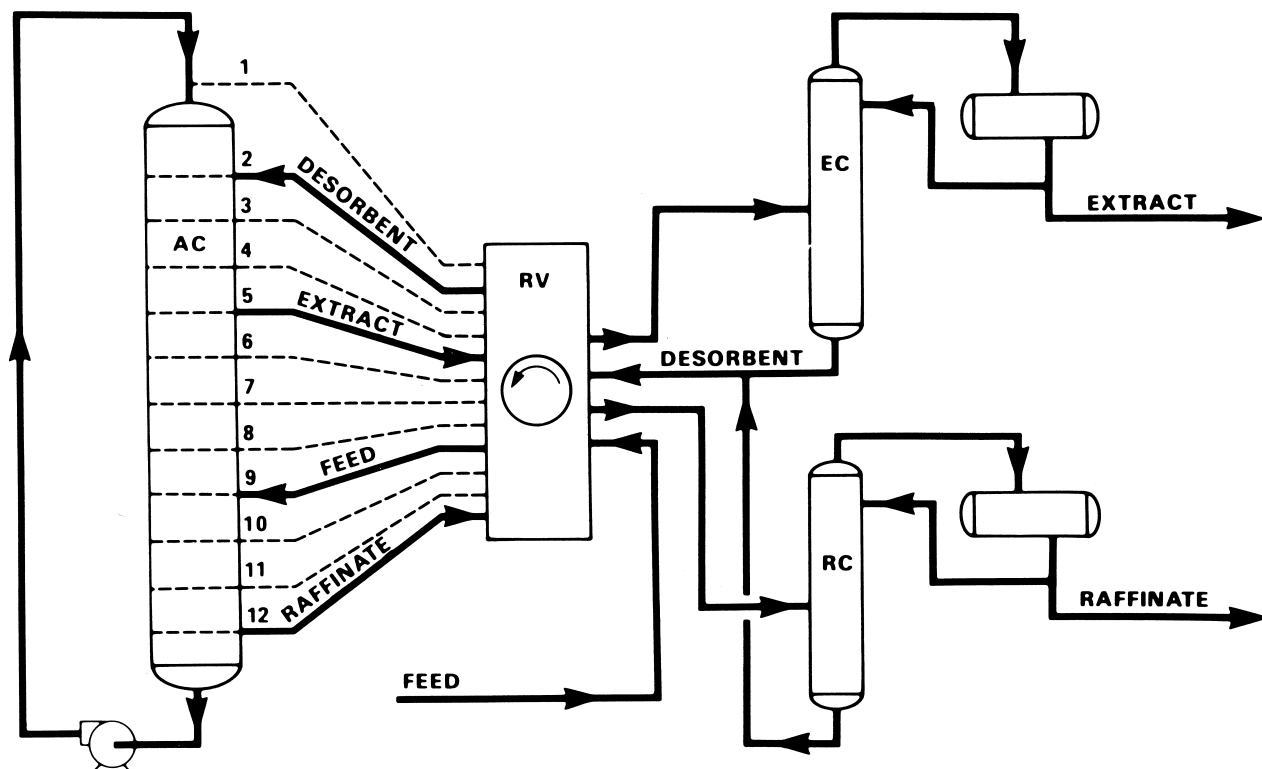


FIGURE 14 Schematic diagram of Sorbex simulated countercurrent adsorption separation system. AC, adsorbent chamber; RV, rotary valve; EC, extraction column; RC, raffinate column. (Reprinted with permission of UOP Inc.)

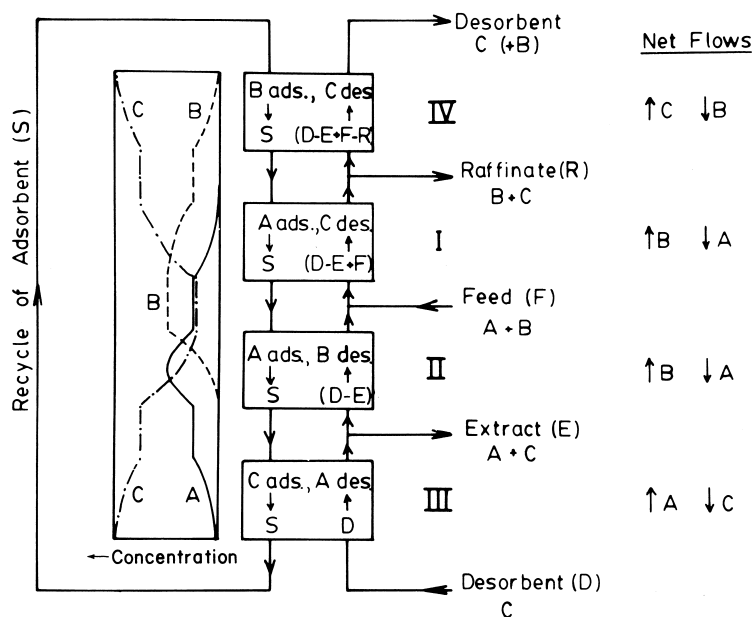


FIGURE 15 Schematic diagram showing the roles played by the four principal sections of a Sorbex system with the required net flow directions. (Reprinted with permission from Ruthven, D. M. (1984). "Principles of Adsorption and Adsorption Processes," copyright John Wiley & Sons, New York.)

TABLE IV Commercial Sorbex Processes^a

Name	Feed	Extract	Raffinate	Process details
Parex	Mixed C ₈ aromatics	98–99% PX	OX, MX, EB	K–BaY + toluene as Sr–BaX + PDEB or K–BaX + PDEB
Ebex	Mixed C ₈ aromatics	OX, MX, PX	99% EB	NaY or Sr–KX + toluene
Molex	<i>n</i> -Alkanes, branched alkanes, and cycloalkanes	<i>n</i> -Paraffins	Branched and cyclic isomers	5A Sieve + light paraffin desorbent
Olex	Olefins + paraffins	Olefins	Mixed paraffins	Probably CaX or SrX
Sarex	Corn syrup	Fructose	Other sugars	Aqueous system CaY

^a Abbreviations: OX, *o*-xylene; MX, *m*-xylene; PX, *p*-xylene; EB, ethylbenzene; PDEB, *p*-diethylbenzene.

where, as a result of the very low concentrations of the contaminants, the LUB is large, so that very large beds would be needed for a conventional cyclic batch process.

B. The Sorbex Process

A more sophisticated development of the same general principle is the Sorbex process, developed by UOP, which is illustrated in Fig. 14. In this system a single fixed adsorbent bed is divided into a number of discrete sections, and the feed, desorbent, raffinate, and extract lines are switched through the bed by a rotary valve. The process operates essentially isothermally with regeneration of the adsorbent by displacement desorption. There are four distinct zones in the bed, with changes in liquid flow rate between zones. Each zone consists of several sections (Fig. 14).

The operation is most easily understood by reference to the equivalent true countercurrent system (Fig. 15). If we consider a feed containing two species A and B, with A the more strongly adsorbed, and a desorbent C, then in order to obtain separation the net flow directions in each section must be as indicated. With the equilibrium isotherms and the feed composition and flow rate specified, this requirement in effect fixes all flow rates throughout the system as well as the adsorbent recirculation rate or switch time. From simple theoretical considerations it can be easily shown that the affinity of the adsorbent for the desorbent should be intermediate between that for the strongly and weakly adsorbed feed compounds (i.e., $\alpha_{AC} > 1.0$, $\alpha_{BC} < 1.0$). The heights of the individualized bed sections are then determined by the requirement that each section contain sufficient “theoretical plates” to achieve the required purity of raffinate and extract products. For a linear system the analysis is straightforward since simple expressions for the concentration profile are available in terms of the kinetic and equilibrium

parameters. The analysis for a nonlinear system is more complicated and requires numerical simulation of the system.

Detailed reviews of the modeling and optimization of such processes have been given by Ruthven and Ching (1989) and by Morbidelli et al. (1989, 1995) (see references given in the bibliography).

Large-scale Sorbex processes have been developed for a variety of different bulk separations; a brief summary is given in Table IV. In recent years, the same principle has been applied also to a wide range of chiral separations and other “difficult” separations that are important in the pharmaceutical industry. Several novel system configurations have been developed. In one system, a carousel of 12 small columns rotates between two stationary circular headers, which act as the switch valve, thus effectively incorporating the adsorption and the flow switching functions within a single unit.

SEE ALSO THE FOLLOWING ARTICLES

● ABSORPTION (CHEMICAL ENGINEERING) ● CHEMICAL THERMODYNAMICS ● CHROMATOGRAPHY ● DISTILLATION ● KINETICS (CHEMISTRY) ● PETROLEUM REFINING ● SOLVENT EXTRACTION ● ZEOLITES, SYNTHESIS AND PROPERTIES

BIBLIOGRAPHY

- Barrer, R. M. (1978). “Zeolites and Clay Minerals as Sorbents and Molecular Sieves,” Academic Press, New York.
- Basmodjian, D. (1997). “The Little Adsorption Book,” CRC Press, Boca Raton, FL.
- Breck, D. W. (1974). “Zeolite Molecular Sieves,” John Wiley & Sons, New York.
- Do, D. D. (1998). “Adsorption Analysis: Kinetics and Equilibria,” Imperial College Press, London.

- Helfferich, F., and Klein, G. (1970). "Multicomponent Chromatography," Marcel Dekker, New York.
- Kärger, J., and Ruthven, D. M. (1992). "Diffusion in Zeolites and other Microporous Solids," John Wiley & Sons, New York.
- Krishna, R., and Wesselingh, J. A. (1997). "The Maxwell-Stefan Approach to Mass Transfer," *Chem. Eng. Sci.* **52**, 861–911.
- Rodrigues, A. E. et al., eds. (1989). "Adsorption: Science and Technology," Kluwer Academic, Dordrecht, Holland.
- Ruthven, D. M. (1984). "Principles of Adsorption and Adsorption Processes," John Wiley & Sons, New York.
- Ruthven, D. M., and Ching, C. B. (1989). "Counter-Current and Simulated Counter-Current Adsorption Separation Processes," *Chem. Eng. Sci.* **44**, 1011–1038.
- Ruthven, D. M., Farooq, S., and Knaebel, K. (1994). "Pressure Swing Adsorption," VCH, Weinheim, New York.
- Suzuki, M. (1990). "Adsorption Engineering," Kodansha-Elsevier, Tokyo.
- Valenzuela, D. P., and Myers, A. L. (1989). "Adsorption Equilibrium Data Handbook," Prentice Hall, Englewood Cliffs, NJ.
- Wakao, N. (1982). "Heat and Mass Transfer in Packed Beds," Gordon & Breach, New York.
- Wankat, P. (1986). "Adsorption Separation Processes," CRC Press, Boca Raton, FL.
- Whyte, T. E., Yon, C. M., and Wagner, E. A., eds. (1983). "Industrial Gas Separations," Am. Chem. Soc., Washington, DC.
- Yang, R. T. (1986). "Gas Separation by Adsorption Processes," Butterworth, Stoneham, MA.



Aerosols

G. M. Hidy

Envair/Aerochem

- I. Phenomenological Aspects
- II. Physical and Chemical Properties
- III. Kinetic Theory of Aerosols
- IV. Production of Aerosols
- V. Measurement Principles
- VI. Industrial Gas Cleaning

GLOSSARY

Brownian motion Thermal agitation of particles resulting from collision of particles with gas molecules.

Coagulation Process of collision and sticking of particles resulting in agglomeration, an increase in effective particle size, and a reduction in concentration of suspended particles.

Diffusion Random migration of particles in a favored direction resulting from Brownian motion or turbulent eddy motion of the suspending gas.

Impaction Collision of particles on an obstacle as a result of the action of inertial and viscous forces acting on the particle.

Interception Collision of particles with an obstacle resulting from aerosol flow and the finite size of particles.

Light extinction Loss of light from a pathway from scattering and absorption of light by particles and gas molecules.

Nucleation Process of formation of new particles from a supersaturated vapor or a chemical reactive gas.

Phoretic forces Forces on suspended particles resulting from differential molecular collisions on the par-

ticle surface or differential, incident electromagnetic radiation.

Size distribution Distribution of particle concentration with particle size.

AEROSOLS, aerocolloids, or aerodispersed systems are collections of tiny particles suspended in gases. They include clouds of suspended matter ranging from haze and smoke to dusts and mists, fogs, or sprays. The science and technology of aerosols matured rapidly in the twentieth century as a result of the increasing interest in their chemistry and physics.

Aerosols vary widely in properties depending on the nature of the suspended particles, their concentration in the gas, their size and shape, and the spatial homogeneity of dispersion. The term is generally restricted to clouds of particles that do not readily settle out by gravity, creating a stable suspension for an extended period of time. They exist in nature as part of planetary atmospheres. Aerosols have extensive involvement in technology, ranging from agricultural sprays to combustion, the production of composite materials and microprocessor technology. They are of concern because of their contribution to hazards in the

workplace and air pollution. They are sometimes hazardous as explosive mixtures.

Both liquid and solid material can be suspended in a gas by a variety of mechanisms. Aerosols produced under laboratory conditions or by specific generating devices may have very uniform properties that can be investigated relatively easily by physical and chemical instrumentation. Natural aerosols found in the atmosphere are mixtures of materials from many sources that are highly heterogeneous in composition and physical properties. Their characterization has required the application of a variety of measurement techniques and has been a major activity in modern aerosol science.

I. PHENOMENOLOGICAL ASPECTS

A. Classification

1. Dusts

Dusts are clouds of solid particles brought about by mechanical disintegration of material, which is suspended by mixing in a gas. Examples include clouds of particles from the breakup of solids in crushing, grinding, or explosive disintegration and the disaggregation of powders by air blasts. Dust clouds are often dramatic in form as storms rising from the earth's surface and traveling hundreds of miles. Generally, dusts are quite heterogeneous in composition and have poor colloidal stability with respect to gravitational settling because they are generally made up of large particles. Yet, the lower range of their particle size distribution may typically be submicroscopic.

2. Smokes

In contrast to dusts, smokes cover a wide variety of aerial dispersions dominated by residual material from burning, other chemical reactions, or condensation of supersaturated vapors. Such clouds generally consist of smaller particles than dusts and are composed of material of low volatility in relatively high concentrations. Because of the small size of the particles, smokes are more stable to gravitational settling than dusts and may remain suspended for an extended period of time. Examples include particulate plumes from combustion processes, chemical reactions between reaction gases such as ammonia and hydrogen chloride or ozone and hydrocarbon vapors, oxidation in a metallic arc, and the photochemical decomposition of materials such as iron carbonyl. An important measure of smoke is particle size; the distribution in size is constrained to be smaller than 10 micrometers (μm) in diameter to less than a tenth of a μm . Smokes normally have high concentrations, often exceeding 10^4 particles/ cm^3 . In the atmosphere, smoke from chimneys obscuring vis-

ibility is a common sight. In most modern cities, smoke plumes have largely been eliminated with pollution.

When smoke formation accompanies traces of noxious vapors, it may be called a fume—for example, a metallic oxide developing with sulfur in a melting or smelting process. The term *fume* is also used in a more general way to describe a particle cloud resulting from mixing and chemical reactions of vapors diffusing from the surface of a pool of liquid.

3. Mists

Suspensions of liquid droplets by atomization or vapor condensation are called mists. These aerial suspensions often consist of particles larger than $1\ \mu\text{m}$ in diameter, and relatively low concentrations are involved. With evaporation of the droplets or particle formation by condensation of a vapor, higher concentrations of very small particles in the submicrometer size range may be observed. In general, mists refer principally to large-particle suspensions such that historically particle size is the principal property distinguishing mists from smokes. If the mist has sufficiently high particle concentration to obscure visibility, it may be called fog. Hazes in the atmosphere usually contain relatively high concentrations of very small particles with absorbed liquid water. The name *smog* (*smoke* combined with *fog*) refers to a particulate cloud normally observed over urban areas, where pollutants mix with haze and react chemically to contribute condensed material to the particulate mixture.

4. Colloidal Stability

The term *aerosol* has been associated with F. G. Donnan in connection with his work on smokes during World War I. An aerosol is regarded as an analogy to a liquid colloidal suspension, sometimes called a hydrosol. These suspensions are relatively stable, with very low gravitational settling speeds and slow rates of coagulation. The stability to gravitational settling is the principal criterion for defining an aerosol.

Low settling rate in itself is not adequate for defining an aerosol. Additional criteria have emerged. For example, the thermal agitation or Brownian motion of particles is an important characteristic of aerosol particles. Brownian motion becomes a factor for particle behavior of particles less than $0.5\ \mu\text{m}$ in diameter. Brownian motion essentially provides the theoretical linkage between the idealized behavior of molecules and small particles. The mechanical theory of large molecules and spheres in gases applies well to the behavior of very small particles in the submicroscopic size range. This characterization is central to the evolution of a large segment of particle science and technology. Indeed, it forms the basis for explaining features of

cloud behavior, particle sampling, the description of their depositional behavior, and their removal from industrial gas streams.

B. Natural Phenomena

Aerosols are readily observed in nature. The atmospheres of planets of the solar system are rich in suspended particulate matter, as in interplanetary and interstellar space. The wealth of visual experience in observing the planets depends on gases and particles concentrated in their atmospheres. The variety of color and opacity of atmospheres is a direct result of light absorption and scattering from particles as well as their suspending gases. Individual particle clouds are frequently identifiable in planetary atmospheres. They show the broad features of atmospheric motion as giant swirls, veils, streaks, and puffs. The best known planetary aerosols are those of the earth. The earth's atmosphere is rich in suspended particles. Their presence has been observed and reported in the literature for centuries. Yet only since the early 1960s has scientific instrumentation become available to characterize atmospheric aerosols in great detail.

Airborne particles in the earth's atmosphere probably were recognized first in relation to sea spray drift or dramatic events associated with volcanic eruptions and forest or brush fires. However, the haze associated with sea spray and blowing soil or pollen dusts also contributes large quantities of particulate material to the atmosphere. Only in recent years has the significance of the contribution to the earth's air burden of extraterrestrial dust and the *in situ* production of particles by atmospheric chemical reactions become known. The latter is of particular interest in that the oxidation products of sulfurous and nitrogenous gases and certain hydrocarbon vapors are prolific producers of small particles. Thus, the "breathing" of traces of gases from natural biological chemistry in soils such as hydrogen sulfide or ammonia, and pinene or similar vapors from vegetation, actually contributes substantially to the atmospheric aerosol content. The direct transfer of particles to the air is often called *primary emissions*. The materials produced from atmospheric chemical processes are termed *secondary* contributions.

Added to the natural aerosol-forming processes are the emissions from human activities. With the industrialization and urbanization of increasingly large geographic areas, substantial quantities of particulate matter are emitted. The expansion of agriculture has also enhanced the suspension of dust either directly by cultivation or indirectly by deforestation and temporary overproduction, resulting in soil erosion. Pollutant gases, including sulfur dioxide, nitrogen oxides, and certain reactive volatile organic compound (VOC) vapors, also represent substantial potential for particle production in the air.

The estimated rate of particle injection into the air, which characterizes the global aerosol burden, is given in [Table I](#). This table represents a compilation from investigators who have tried to estimate the relative contributions to the atmospheric aerosol. From this survey, the natural contributions far exceed emissions from human activities on a global basis, but locally this is undoubtedly reversed, especially in parts of North America and Europe. From the table, the "best estimate" suggests that about 13% originate with human activity, while the remainder is assigned to natural sources. The importance of particles from atmospheric chemical reactions of gases is also shown from data in the table. More than 13% of the estimated particle burden comes from the secondary processes. Noting that the rates are dominated by large particles in soil dust and sea salt, the secondary fraction is much more important if these sources are not considered. In addition, it readily can be seen that the secondary material should be dominated by particulate sulfur, present as sulfate, and perhaps organic carbon on the basis of these estimates. Indeed, sulfate is a universal constituent of atmospheric particle populations as is carbon.

The enormous quantities of particles injected into the earth's atmosphere are mixed and aged by processes in the air to create a very diverse and complicated mixture. The mix varies greatly with geographic region and with altitude, but also has some remarkably common physical and chemical features. The presence of suspended particles in the earth's atmosphere provides for a variety of natural phenomena and represents an important part of aerosol science. Particulate matter in the air exerts an influence on the transfer of electromagnetic radiation through the atmosphere. This manifests itself in changes in visibility and coloration as a result of light scattering and absorption. A wealth of sky color, shadow, and haziness, which provides a varied and often beautiful setting both for natural objects and for architecture, is a direct result of the influence of suspended particles interacting with visible light.

Changes in the transfer of radiation in different layers of the atmosphere are the crux of the atmospheric energy storage process. Aerosol particles also play a role in distributing solar energy throughout the atmosphere and consequently in affecting climate. A distinctly different function of aerosol particles in the atmosphere involves the formation of clouds of condensed water. Suspended particles basically provide the nuclei for the condensation of moisture and for the nucleation of ice crystals in supercooled clouds. Thus, in a sense, aerosols provide a skeleton through which are derived, with water vapor, rain clouds and precipitation. The opportunity then presents itself for both weather and climate modification by injection of particulate matter into the air.

The interaction between aerosol particles and clouds recently has led to an important theory about the

TABLE I Recent Estimates of Rate (Tg/yr) at which Aerosol Particles of Radius Less than about 20–30 μm Are Produced in, or Emitted into, the Atmosphere^a

Source	“Best” estimate	Range
<i>Natural particles</i>		
Soil and rock debris	1500	60–2000
Forest fires and slash burning debris	50	50–1500
Sea salt	1300	1000–10,000
Volcanic debris	33	15–90
Gas-to-particle conversion in the atmosphere		
Sulfate from sulfur gases	102	130–300
Ammonium salts from ammonia	—	80–270
Nitrate from nitrogen oxides	22	22–300
Organic carbon from plant VOC exhalation	55	55–1000
Subtotal	3062	1410–15,500
<i>Anthropogenic particles</i>		
Particles from direct emissions (combustion, industry, etc.)	120	10–120
Gas-to-particle conversion in the atmosphere		
Sulfate from sulfur dioxide	140	130–200
Nitrate from nitrogen oxides	36	30–36
Organic carbon from VOC emissions	90	15–90
Subtotal	386	185–446
Total	3450	1600–15,900
<i>Extraterrestrial dusts</i>	10	0.1–50

^a Composite of post-1971 estimates. [Adapted from Wolf and Hidy (1999). *J. Geophys. Res.* **102**, 11-113–11-121.]

surprising depletion of ozone in the high atmosphere, the stratosphere. In 1985, English scientists reported a broad region of springtime reduction in stratospheric ozone concentration over Antarctica at an altitude range between 10 and 20 km. This widespread depletion of the stratospheric ozone layer has been named the “ozone hole” in the popular media. The observations were inconsistent with expectations of the gas-phase photochemistry of chlorine, which appears to originate mostly from manmade halocarbons, such as freon refrigerants, rising from the earth’s surface. In the 1990s, scientists postulated that the polar stratospheric clouds made up of sulfuric acid, nitric acid, and water at very cold temperatures, combined with sunlight, provided a medium for the ozone-depleting reactions. Photochemical reactions of chlorine compounds on the ice-aerosol particle surfaces provide for production of chlorine atoms, which in turn interfere with the photochemical ozone cycle in the stratosphere to create the depletion phenomenon.

C. Particle Technology

Aerosol science has found its way into a wide variety of technological applications. Perhaps best known is the use of spray generation principles for manufacturing dis-

persable consumer products such as personal deodorants, household sprays and cleaners, and pesticides. The use of aerosol technology is widespread in agriculture for the dispersal of pesticides. There is an extensive application in the field of fine-particle production and the use of these particles for material surface coatings, reinforcement and strengthening of composites, and production of microelectronic chips and components. An obvious application also enters into the engineering of fuel combustion systems.

The concise scientific definition of an aerosol refers specifically to a colloidal state of material suspended in a gas. However, the term has acquired an additional meaning in common household usage. In the commercial packaging field, the term *aerosol* now is synonymous with pressurized products that are released in a dispersed form from a can or a bottle. The discharge ranges from coarse fogs and mists to finely divided liquid or powder dispersions.

Although the list of products that can be dispersed by the aerosol method is extensive, they have common characteristics. The materials are packaged under pressure and are released by pressing a simple valve. They contain an active ingredient and a propellant that provides the force for expelling and breaking up the product. In many cases, the carrier or solvent for the active ingredient is included in the suspension to make a useful product formulation.

The use of devices to disperse quantities of pesticides for agricultural or public health applications has been widespread over the world. Their application ranges from individual household and domestic activity to very large-scale, systematic treatment as an integral part of agricultural practice.

In general, the control of pests or disease involves the distribution of a small amount of pesticide over a very large surface area. This may include surfaces of buildings, vegetation, or soil. The dispersal of pesticides is accomplished by suspending material in a liquid, usually water, and then spraying or by dusting with a finely divided powder. A variety of sprayers are available for dispersing pesticides; techniques have been developed for different applications, involving ground-based or aircraft operations.

Some optimum droplet size range is recognized as the most effective for each pesticide and for each formulation used for a specific control problem. Maximum effective control of a target organism with minimum use of toxic materials and minimum adverse impact on the surrounding ecosystem is the objective. This simple statement covers a highly complex physical and biological phenomenon that occurs during and after an area application of pesticides. Research toward this objective has been conducted for many years. The earliest work with Paris green and toxic botanicals progressed through petrochemical products, culminating in the extensive use of the synthetic pesticides, organochlorines, as well as organophosphorus and carbamate materials.

In some applications of aerosol technology, the capability of generating aerosols with large volumes of suspended material has been developed. Some situations dictate uncontrolled smoke dispersal, as in military applications. However, others can involve at least a degree of control of particle diameter and the integrity of chemical composition of the aerial suspension. Another application of control requirements for sprays and mists is in the area of medical research and therapy. For example, aerosols have been used for therapeutic treatment of respiratory disease. Here, the medication must be dispersed with a controlled particle size and volume for a long period of time under conditions when any chemical change in the suspended material is negligible. Requirements for the study of the influence of air contaminants on respiratory disease led to the engineering of large exposure chambers with carefully controlled air properties. The controlled environment and clinical conditions can be applied equally well to clean room environments. The latter is an important adjunct to particle control technology in modern industry, where manufacturing requires very sterile conditions.

The dispersal of material by spraying or by dusting of solid particles plays an important role in combustion technology. Large industrial boilers or furnaces employ oil fuel

or pulverized coal injection to provide an inlet stream for efficient combustion. Diesel engines, turbines, and certain kinds of rocket engines use fuel spray injection. The process of combustion concerns the steps of transport of fuel and oxidizer to the reaction or flame zone. Because of its technological importance, considerable research has been done on the burning of finely divided particles.

The main design objectives for combustion devices using finely divided fuels are the following:

1. A high combustion intensity
2. A high combustion efficiency so that as little unreacted fuel leaves the combustion chamber as possible
3. A stable flame
4. The minimum deposition of soot or solid on the combustion chamber walls
5. The maximum rate of heat transfer from the flame to a heat sink or exchanger

The most common method of firing liquid fuels is to atomize the liquid before combustion. The fuel is introduced into the combustion chamber in the form of a spray of droplets, which has a controlled size and a velocity distribution. The main purpose of atomization is to increase the surface area of the liquid in order to intensify vaporization, to obtain good distribution of the fuel in the chamber, and to ensure easier access of the oxidant to the vapor. Injection of powdered fuels follows similar principles.

After atomization, combustion takes place through a series of processes, the most important of which include the following:

1. Mixing of the fuel particles with air and hot combustion products, a process usually occurring under turbulent conditions
2. Transfer of heat to the particles by convection from the preheated oxidant and recycled combustion gases and by radiation from the flame and the chamber walls
3. Evaporation of particles; often accompanied by cracking of vapor
4. Mixing of the vapor with air and combustion gases to form an inflammable mixture
5. Ignition of the gaseous mixture (depending on the mixing conditions, an ignition may occur at the oxygen-rich boundaries of eddies containing many vaporizing particles or may occur as a microscale process surrounding an individual particle)
6. Formation of soot, with residual fuels
7. Combustion of soot, a relatively slow process

In practice, these processes often occur simultaneously, resulting in an extremely complex aerosol system. Owing

to this complexity, research has been centered on those areas considered to be the most important for practical applications.

A combustion aerosol differs from a premixed, combustible gaseous system in that it is not uniform in composition. The fuel is present in the form of discrete particles, which may have a range of sizes and may move in different directions with different velocities than the main stream of gas. This lack of uniformity in the unburned mixture results in irregularities in the propagation of the flame through the spray and, thus, the combustion zone is geometrically poorly defined.

The process of particle combustion can be illustrated by the simplest case, that of a one-dimensional laminar flame moving at low velocity. The flame can be considered to be essentially a flowing reaction system in which the time scale of the usual reaction rate expression is replaced by a distance scale. As the unburned particle approaches the flame front, it first passes through a region of preheating, during which some vaporization occurs. As the flame zone is reached, the temperature rapidly rises and the particle burns. The flame zone can thus be considered a localized reaction zone sandwiched between a cold mixture of fuel and oxidant on one side and hot burned gases on the other; if the gas flow through the flame is one-dimensional, the flame front is planar. The nature of the burned products depends on the properties of the spray in its unburned state. If the particles are large, combustion may not be complete in the main reaction zone and unburned fuel penetrates well into the burned gas reaction. If the particles are small, a state of affairs exists that approximates very closely the combustion of premixed gaseous flame. Here, the particles are vaporized in the preheat zone, and reaction after that is between the reactants in their gaseous state. The other factors that determine the time to reach complete combustion (that is, the length of the combustion zone) are the volatility of the liquid fuel, the ratio of fuel to oxidant in the unburned mixture, and the uniformity of the mixture.

In most practical systems such as a furnace or a rocket engine, the combustion process is much more complicated due to two important factors. First, to a large extent the mixing of the fuel and oxidant takes place in the combustion chamber and thus the mechanics of the mixing process plays an important role. Second, the flow patterns are complicated by turbulence or recirculation and frequently cannot be represented by simplified theories.

A fair number of dusts are capable of sustaining flames; the number exceeds more than 100. However, the one of principal technological importance is coal dust. The last quarter of the twentieth century saw a major resurgence of coal as a stable energy source in many nations. Because of the large capital investment in coal-fired systems, the pressures of air pollution emission control, and interest in

coal-based synthetic fuels, research on coal combustion has surged ahead. As with fuel sprays, a modern description of the particle burning process is complicated by the interactions of diffusion and chemical kinetics.

The process of particle combustion depends on the physical and chemical nature of the solid as it heats and burns. Coal is a complex material of volatile and non-volatile components which becomes increasingly porous during volatilization of low-boiling constituents in burning. The crucial practical questions for boiler design concern whether pulverized fuel combustion is controlled by oxidizer diffusion or by chemical kinetics.

An important by-product of the combustion of liquid or vapors is soot particles. Carbonaceous particle production is also a serious limitation in the use of diesel engine technology for transportation, because of air pollution. However, the limitation is of "benefit" in another industry. The production of carbon black is a major industry, with widespread use of the product for binders and dyes.

The use of fine powders for industrial applications has become an increasingly important factor in aerosol technology. Finely divided powders now are used for the reinforcement of materials, surface coatings, and laminated, polycomponent materials.

One of the more interesting applications of aerosol particle technology is in the rapidly expanding field of microelectronics. The production of electronically active surface films on substrates by the deposition of semiconductor particles is a rapidly advancing technology. A potentially important extension of this technology is the production of nanoparticles in the 0.01- μm diameter range. Surface films imbedding these particles can have unique microelectronic properties that offer opportunities for making molecular-level semiconductor "quantum" dots or wires imbedded in other semiconductors, or "quantum well pyramids" that have special luminescence or optical properties of practical interest.

D. Environmental Influence

Aerosols have an adverse effect on human health and create hazards to public safety. Particle suspensions are involved in catalyzing respiratory disease in the workplace and home or through pollution of ambient air. In the United States, other environmental disturbances, including potential effects on biota, accelerated material deterioration, and visibility degradation, are attributed to pollution aerosols. Particle suspensions also can affect safety because of their potential for inflammability and explosion. Dust explosions have occurred in a variety of industrial situations including grain storage and manufacturing areas where particle suspensions are produced. The hazards of toxic dust and fume release are well documented in the work-place,

and considerable effort must be exercised sometimes to prevent worker exposure to them.

Public concern for the hazards of particle suspensions in the indoor and outdoor environment has produced regulations limiting particle concentrations and exposure levels. In the workplace, dust hazards are constrained by total mass concentration as well as concentration of specific toxic chemicals. In the ambient air, protection is stipulated in terms of total mass concentration of suspended particles and certain chemical species, namely, lead and sulfate. Recently, measures of exposure have begun to distinguish between fine particles less than $2.5\ \mu\text{m}$ and coarse particles between 2.5 and $10\ \mu\text{m}$. This separation relates to the ability of particles to penetrate the human respiratory system, and to different sources of fine and coarse particles.

One of the most common airborne suspensions known to affect the respiratory system adversely is tobacco smoke. The chemicals in tobacco smoke include a number of carcinogens, including nicotine and some of its derivatives, as well as poisonous gases including carbon monoxide and nitrogen dioxide.

II. PHYSICAL AND CHEMICAL PROPERTIES

The gases in which particles are suspended retain their normal physical and chemical properties, taking into account exchange processes with the particles. The suspended particles have properties that correspond to condensed material. These include their surface, volume, mass, mass density, surface energy, freezing point (if liquid), heat of vaporization or sublimation, solubility, heat of adsorption for gases, vapor pressure, viscosity or elastic properties, thermal conductivity, diffusivity of components, magnetic and electric properties, dielectric constant index of refraction, chemical reactivity, radioactivity, and momentum and energy properties.

A. Critical Physical Properties

Perhaps the properties most critical physically to aerosol particles are those related to size and distribution of size in a cloud.

It is generally assumed that the substances making up the aerosol particle possess the same characteristics as the substance in macroscopic amounts, which is termed the *bulk state*. Then the various physical properties of the particle, such as those mentioned above, are either known or easily determined by standard techniques. These techniques generally do not involve direct measurements on aerosol particles. The assumption of bulk state behavior, however, may not be justified generally for small aerosol

particles. The physical properties of small particles of a given substance can be quite different from the corresponding properties of the same substance in the bulk state.

Deviations in behavior of small aerosol particles from the bulk state are widely recognized for many physical properties such as vapor pressure, freezing point, and crystal structure. Yet there has been a lack of a systematic classification of these deviations. Also, although deviations are expected to occur for small particles, there have been few experimental measurements of such deviations owing partly to the difficulties of making such measurements.

We classify a deviation from the bulk properties of the properties of small aerosols as either extrinsic or intrinsic. Extrinsic deviations are associated with characteristics of particles that are not inherent but are caused by external agents such as the mode of formation of the particle or the absence of phase-transition nuclei in the particles. Thus, extrinsic deviations are associated more with a lack of control in the particle generation process than with any fundamental cause. Intrinsic deviations may occur in several ways. One type of intrinsic deviation is associated with the radius of curvature of small particles. For example, it is well known that a liquid droplet with a given radius has a higher vapor pressure than that found with another droplet of the same composition but of larger radius. For sufficiently small particles, another type of intrinsic deviation may occur. Here, the intermolecular energy of interaction of molecules making up a very small particle is altered by the fact that a given molecule does not interact with an extremely large number of other molecules, but instead can interact only with a limited number within the particle. Furthermore, if the aerosol particle is still smaller, almost molecular size of order $\lesssim 0.01\ \mu\text{m}$, in the “nanoparticle” range, molecular fluctuations will be so large that it is probably no longer meaningful to speak of the physical properties of a particle of such small size in the usual macroscopic sense. Examples of “near-molecular” behavior in nano-particle formed ceramics is super plastic behavior. Band gap energy levels in semiconductors are also increased by quantum effects.

One type of extrinsic deviation is found in the lowering of the freezing point or the raising of the boiling point for small liquid droplets from that for the bulk state. Such effects are usually attributed to the absence of phase transition nuclei. The absence of such nuclei stems from the fact that the bulk material from which the aerosol particles are formed probably contains only minute traces of foreign material (nuclei) per unit volume, so that there is only a very small probability that any small aerosol particle will contain even one nucleus. This circumstance results in the situation that nearly all aerosol particles formed by vapor condensation and subsequent cooling well below the melting point of the parent material are likely to be in a

metastable liquid state. For example, sodium chloride has been observed to exist as a relatively stable subcooled liquid at room temperature, hundreds of degrees below the melting point of crystalline sodium chloride.

Another example of an extrinsic deviation was found by examining absorption spectra from freshly formed aerosols composed of iron carbonyl, 30–200 Å in size. Among the spectra were some corresponding to excited states of carbon monoxide, as well as bands that were possibly associated with a molecular oxygen transition. The oxygen excitation had energy levels of 7–9 eV, suggesting that the excitation was not due to chemical reactions or incident photons only. It is possible that the spectral absorption was also related to gas molecules adsorbed on the iron surface or to large surface energy of the small particles. When the small particles coagulate or surface crystallites are relocated, a large amount of energy may be released and transferred to gas molecules adsorbed on the particle surface. In this way, certain high excitation levels may be populated in a manner differing from that predicted by thermal equilibrium.

Other types of extrinsic deviation are found in the special properties imparted to small particles by the manner of their preparation. For example, production of small particles by grinding in a mill alters the heat of adsorption of gases by the particles. In general, the method of particle production may introduce defects or microscopic impurities that differ from what is found in the parent material. Often extrinsic and intrinsic deviations occur in the same physical property, as in the example of supercooling of sodium chloride cited above. This fact makes the study of intrinsic deviations very difficult.

Intrinsic deviations are perhaps most widely known through the effect of the radius of curvature of small particles on many physical properties such as vapor pressure, freezing point, surface tension, heat of evaporation, and others. Intrinsic deviations not directly associated with the radius of curvature have been observed by X-ray crystallographic studies of very small crystallites with radii less than 0.01 μm . In these studies, the lattice spacings observed in the small crystallites differed significantly from the lattice spacings observed for the bulk state of the parent material. The effect of such alterations on various physical properties has not been studied. In general, one expects that for particles of radius less than $\sim 0.01 \mu\text{m}$, intrinsic deviations of this sort must occur; however, it has been obviously very difficult to observe such deviations experimentally. Only recently has substantial progress been made in characterizing unique properties associated with the nanoparticle regime.

Another type of intrinsic property is derived from the theory of light scattering in particles. The phenomenon of Raman and fluorescent scattering from molecules suspended in small dielectric particles exemplifies such prop-

erties. Scattered light is affected by the morphology and optical properties of the particle and the distribution of optically active molecules within it. The light scattered and its angular distribution are quite different from those found when the molecules are distributed within the same material in bulk.

1. Cohesive and Adhesive Forces

Particles in dusts or powders tend to stick together remarkably well. The suspension of powders depends critically on the agglomeration characteristics. Once suspended, the capability of particles to agglomerate after collision also depends on the attractive forces of interaction after contact. It is difficult to break up aggregates of particles to produce clouds of nonagglomerated material. The capacity of particles to stick together indiscriminately is the result of weak attractive forces between molecules as well as bipolar electrostatic forces. These forces have been named cohesive and adhesive, depending on the heterogeneity of material at the boundary between particles. The distinction between cohesion and adhesion in the literature on fine powders is somewhat fuzzy, but we shall adopt the following conventions, which are consistent with classical definitions in physics. *Cohesion* is the tendency for parts of a body of like composition to hold together. This implies that cohesive forces arise between like molecules in a solid or between small particles of the same composition. *Adhesion*, on the other hand, refers to attraction across the boundary or interface between two dissimilar materials. Thus, adhesive forces are likely to be the most common attractive forces in all but artificially generated aerosols.

B. Particle Size Distribution

In practice, aerosols in nature and in technology cover a broader size range than called for by their rigorous scientific definition. Normally, the range of interest is less than 100 μm in diameter, extending to molecular dimensions. A summary of particle dispersoids, methods of measurement, gas cleaning equipment, and mechanical parameters is given in Fig. 1. A striking and important feature of aerosols is illustrated in the figure. Particles range over five orders of magnitude in size. By analogy, this roughly corresponds to a domain from sand grains to tall buildings in a city. Thus, the microscopic world of fine-particle suspensions should be as diverse and rich as our everyday macroscopic environment. This range poses major challenges to the scientist for developing theory, for measurement, and for mechanical production and removal.

The theory of particle clouds proceeds from consideration of the dynamics of the particle size distribution function or its integral moments. This distribution can take two forms. The first is a discrete function in which particle

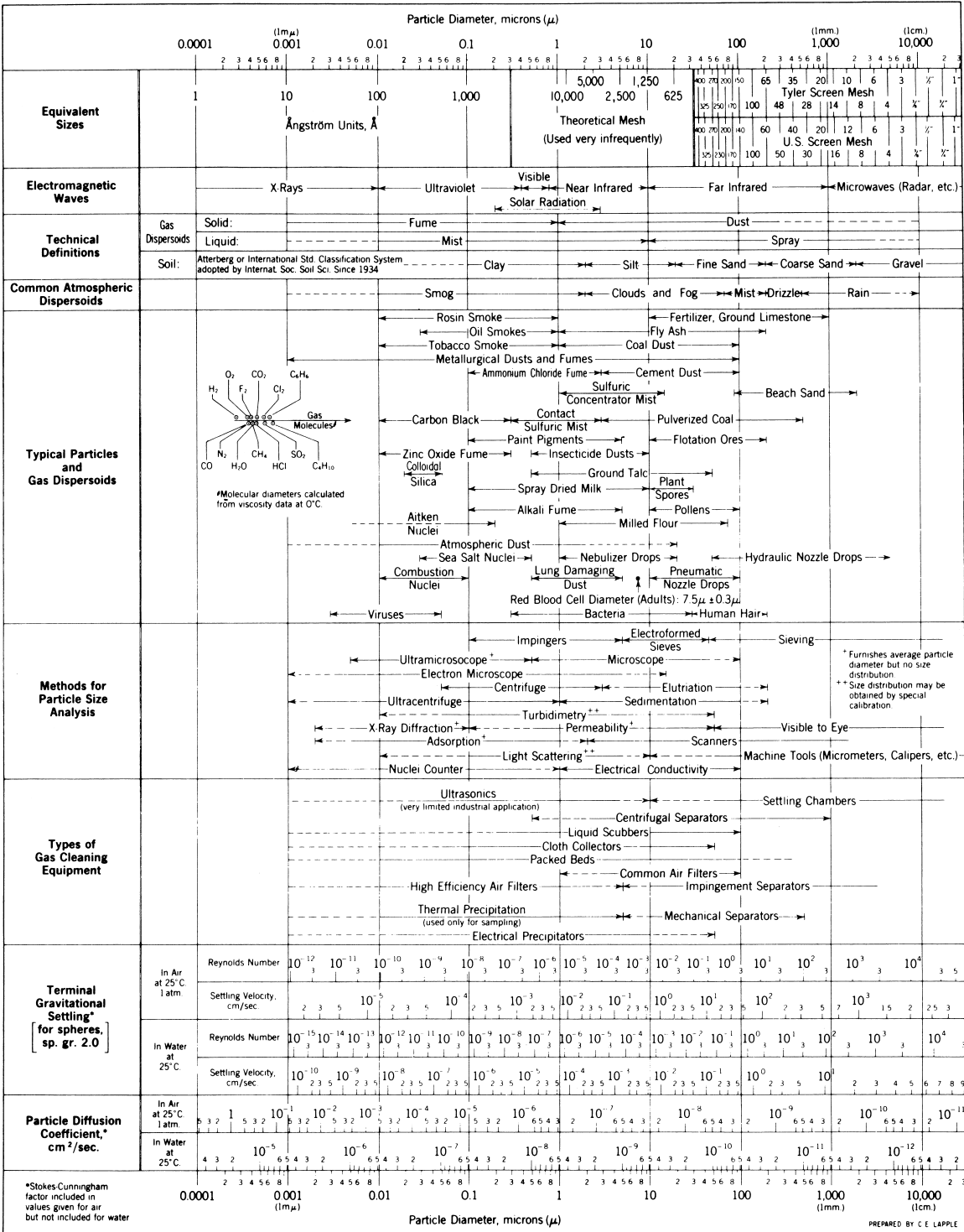


FIGURE 1 Summary diagram of particle properties as a function of size, including measurements and gas-cleaning technology. [Courtesy of Stanford Research Institute.]

sizes can only be multiples of a singular species. As an example, consider the coagulation of a cloud of particles initially of a unit size. Then, after a time, all subsequent particles will be k th aggregates of the single particle, where $k = 1, 2, 3, \dots$ represents the number of unit particles per aggregate. Physically, the discrete size distribution is appealing since it describes well the nature of the particulate cloud. The second function, continuous distribution, is usually a more useful concept in practice. This function is defined in terms of the differential dN , equal to the number of particles per unit volume of gas at a given point in space ($\mathbf{r}$) at time t in the particle volume range, $\mathfrak{v}$ and $\mathfrak{v} + d\mathfrak{v}$. The distribution function then is

$$dN = n(\mathfrak{v}, \mathbf{r}, t) d\mathfrak{v}$$

Although this form accounts for the distribution of particles of arbitrary shape, the theory is well developed for spheres. In this case, one can also define the distribution function in terms of the particle radius (or diameter),

$$dN = n_R(R, \mathbf{r}, t) dR$$

where dN is the concentration of particles in size range R and $R + dR$, n_R is the distribution function in terms of radius, and t is time. The radius may be geometric, or it may be used as an aerodynamic or other physical equivalent. An aerodynamic radius is defined in terms of geometric size and particle density, which govern the motion of the particle. Other physical parameters are the optical equivalent radius, which depends on the light-scattering cross section of the particle.

The volume and radius distribution functions are not equal, but can be related by the equation:

$$n_R = 2\pi R^2 n$$

The moments of the size distribution function are useful parameters. These have the form:

$$M(r, t) = \int_0^\infty n_R R^r dR$$

The zeroth moment ($a = 0$),

$$M_0 = \int_0^\infty n_R dR = N$$

represents the total number concentration of particles at a given point and time. The first moment normalized by the zeroth moment gives a number average particle radius:

$$M_1/M_0 = \bar{R} = \int_0^\infty n_R R dR / \int_0^\infty n_R dR$$

The third moment is proportional to the total volume concentration of particles, or

$$V = \frac{4}{3}\pi M_3 = \frac{4}{3}\pi \int_0^\infty n_R R^3 dR$$

where V is the volume fraction of particles in cubic centimeters of material from cubic centimeters of gas. If the particle density is uniform, the average particle volume is

$$\bar{v} = \frac{V}{N} = \frac{4}{3}\pi \frac{M_3}{M_0}$$

The volume mean radius is $3M_1/4\pi M_3$.

In general, particle distributions are broad in size-concentration range so that they often are displayed on a logarithmic scale. For example, data are frequently reported as $\log n_R$ versus $\log R$. Another display is $dV/d \log R$ versus $\log R$. The area under the distribution curve plotted in this way is proportional to the mass concentration of (constant density) particles over a given size range, independent of size.

The shape of the size distribution function for aerosol particles is often broad enough that distinct parts of the function make dominant contributions to various moments. This concept is useful for certain kinds of practical approximations. In the case of atmospheric aerosols; the number distribution is heavily influenced by the radius range of 0.005–0.1 μm , but the surface area and volume fraction, respectively, are dominated by the range 0.1–1.0 μm and larger. The shape of the size distribution is often fit to a logarithmic-normal form. Other common forms are exponential or power law decrease with increasing size.

The cumulative number distribution curve is another useful means of displaying particle data. This function is defined as:

$$N(R, r, t) = \int_0^R n_R(R, r, t) dR$$

It corresponds to the number of particles less than or equal to the radius R . Since $n_R = dN(R)/dR$ the distribution function can be calculated in principle by differentiating the cumulative function.

C. Chemical Properties

The chemical properties of particles are assumed to correspond to thermodynamic relationships for pure and multicomponent materials. Surface properties may be influenced by microscopic distortions or by molecular layers. Chemical composition as a function of size is a crucial concept, as noted above. Formally the chemical composition can be written in terms of a generalized distribution function. For this case, dN is now the number of particles per unit volume of gas containing molar quantities of each chemical species in the range between $\tilde{n}_i$ and $\tilde{n}_i + d\tilde{n}_i$, with $i = 1, 2, \dots, k$, where k is the total number of chemical species. Assume that the chemical composition is distributed continuously in each size range. The full size-composition probability density function is

$$dN = N_g(\Psi, \tilde{n}_2 \cdots \tilde{n}_k, \mathbf{r}, t) d\Psi d\tilde{n}_2 \cdots d\tilde{n}_k$$

Here, $\tilde{n}_i$, the number of moles of a given species, has been eliminated from the function g by the relation, $\Psi = \sum_i \tilde{n}_i \tilde{v}_i$, where $\tilde{v}_i$ is the partial molar volume of species i and $\mathbf{r}$ is a position vector. Since the integral of dN over all $\tilde{v}$ and $\tilde{n}_i$ is N ,

$$\int \Psi \cdots \int_{\tilde{n}_k} g(\Psi, \tilde{n}_2 \cdots \tilde{n}_k, \mathbf{r}, t) d\Psi d\tilde{n}_2 \cdots d\tilde{n}_k = 1$$

Furthermore, the size distribution function can be retrieved by integration over all chemical species.

$$n(\Psi, \mathbf{r}, t) = N \int_{\tilde{n}_2} \cdots \int_{\tilde{n}_k} g(\Psi, \tilde{n}_2 \cdots \tilde{n}_k, \mathbf{r}, t) d\tilde{n}_2 \cdots d\tilde{n}_k$$

The generalized distribution functions offer a useful means of organizing the theory of aerosol characterization for chemically different species. To date, however, the data have not been sufficiently comprehensive to warrant application of such formalism.

III. KINETIC THEORY OF AEROSOLS

Historically a large segment of work in aerosol science has focused on the motion of particles in fluid medium and on the associated heat and mass transfer to that particle. Recent theory has recognized that significant differences exist in momentum, heat, and mass transfer depending on the continuum nature of the suspending medium. This is normally characterized by the ratio of the mean free path of the gas and the particle radius. This ratio is sometimes called the Knudsen number (Kn). For very small Kn , particles behave as if they are suspended in a continuum medium. For very large Kn , the suspending medium is highly rarified and the particles respond to individual collisions of the suspending gas molecules.

Particle Mechanics: The Gas Kinetic Model

Idealization of particle behavior in a gas medium involves a straightforward application of fluid dynamics.

Mechanical constraints on aerosol particle dynamics can be defined by certain basic parameters. Model particles are treated as smooth, inert, rigid spheres in near thermodynamic equilibrium with their surroundings. The particle concentration is very much less than the gas molecule concentration. The idealization requires that the ratio of the size (radius) of gas molecules (R_g) to that of particles i , R_g/R_i , be less than 1 and the mass ratio, $m_g/m_i \ll 1$. Application of Boltzmann's dynamic equations for aerosol behavior requires further that the length ratios $R_g/\lambda_g \ll 1$

and $R_g/L_g \ll 1$, where λ_g is the mean free path of the gas and L_g a typical length scale of the system, such as a spherical collector diameter or a pipe diameter. The theory can be extended to incorporate electrical effects, as well as a coagulation or sticking capacity and gas-condensed phase interactions.

Virtually all of the mechanical theory of particles emerges from a simplification called the single-particle regime. In this situation, particles are assumed to interact only "instantaneously" in collision; otherwise, they can be assumed to behave as a body moving in a medium of infinite extent.

In general, the exchange of momentum between a gas and a particle involves the interaction of heat and mass transfer processes to the particle. Thus, the forces acting on a particle in a multicomponent gas containing molecular gradients (nonuniform) may be linked with their gradients as well as velocity gradients in the suspending medium. The assumption that Kn approaches zero greatly simplifies the calculation of particle motion in a nonuniform gas. Under such circumstances, momentum transfer, resulting in particle motion, is influenced only by aerodynamic forces associated with surface friction and pressure gradients. In such circumstances, particle motion can be estimated to a good approximation by the classical theory of a viscous fluid medium where Kn is zero. Heat and mass transfer can be considered separately in terms of convective diffusion processes in a low Reynolds number regime ($Re \leq 1$). In cases where noncontinuum effects must be considered ($Kn > 0$), the coupling between particle motion and thermal or molecular gradients as in gas nonuniformities becomes important, and so-called phoretic forces play a role in particle motion, but heat and mass transfer again can be treated somewhat independently. Phoretic forces are associated with temperature, and gas component concentration gradients, or electromagnetic forces.

It is common practice to treat particle motion as the basic dynamic scale for transport processes. This is readily illustrated for particles in steady rectilinear motion.

A. Motion of Particles

1. Stokes' Law and Momentum Transfer

When a spherical particle exists in a stagnant, suspending gas, its velocity can be predicted from viscous fluid theory for the transfer of momentum to the particle. Perhaps no other result has had such wide application to aerosol mechanics as Stokes' (1851) theory for the motion of a solid particle in a stagnant medium. The model estimates that the drag force $\mathcal{D}$ acting on the sphere is

TABLE II Characteristic Transport Properties of Aerosol Particles of Unit Density in Air at 1 atm and 20°C^a

Particle radius (cm)	Mobility B (s/g)	Diffusivity D_p (cm ² /s)	Mean thermal speed $\bar{v}_p$ (cm/s)	Mean free path λ_p (cm)
1×10^{-3}	2.94×10^5	1.19×10^{-5}	4.96×10^{-3}	6.11×10^{-6}
5×10^{-4}	5.96×10^5	2.41×10^{-8}	1.41×10^{-2}	4.34×10^{-6}
1×10^{-4}	3.17×10^6	1.28×10^{-7}	0.157	2.07×10^{-6}
5×10^{-5}	6.71×10^6	2.71×10^{-7}	0.444	1.54×10^{-6}
1×10^{-5}	5.38×10^7	2.17×10^{-6}	4.97	1.12×10^{-6}
5×10^{-6}	1.64×10^8	6.63×10^{-6}	14.9	1.20×10^{-6}
1×10^{-6}	3.26×10^9	1.32×10^{-4}	157	2.14×10^{-6}
5×10^{-7}	1.26×10^{10}	5.09×10^{-4}	443	2.91×10^{-6}
1×10^{-7}	3.08×10^{11}	1.25×10^{-2}	4970	6.39×10^{-6}

^a cgs units.

$$\mathcal{D} = 6\pi\mu_g R U_\infty$$

where U_∞ is the gas velocity far from the particle and μ_g the gas viscosity.

The particle mobility B is defined as $B \equiv U_\infty/\mathcal{D}$. Generally, the particle velocity is given in terms of the product of the mobility and a force F acting externally on the particle, such as a force generated by an electrical field. Under such conditions, the particle motion is called “quasi-stationary.” That is, the fluid particle interactions are slow enough that the particle behaves as if it were in steady motion even if it is accelerated by external forces. Mobility is an important basic particle parameter; its variation with particle size is shown in Table II along with other important parameters described later.

The analogy for transport processes is readily interpreted from Stokes’ theory if we consider the generalization that “forces” or fluxes of a property are proportional to a diffusion coefficient, the surface area of the body, and a gradient in property being transported. In the case of momentum, the transfer rate is related to the frictional and pressure forces on the body. The diffusion coefficient in this case is the kinematic viscosity of the gas ($\nu_g \equiv \mu_g/\rho_g$, where ρ_g is the gas density). The momentum gradient is $\mu_g U_\infty/R$.

If the particles fall through a viscous medium by the influence of gravity, the drag force balances the gravitational force, or:

$$\frac{4}{3}(\rho_p - \rho_g)g\pi R^3 = 6\pi\mu_g R q_s$$

where g is the gravitational force per unit mass. Since $\rho_p \gg \rho_g$, the settling velocity is

$$q_s = \frac{2R^2\rho_p g}{9\mu_g}$$

Thus, the fall velocity is proportional to the cross-sectional area of the particle, and the ratio of its density and the gas viscosity (for values, see Fig. 1). If the particle Reynolds number approaches or exceeds unity, Stokes’ theory must be modified.

In terms of the drag coefficient C_D , the drag force is written:

$$C_D = \frac{2\mathcal{D}}{\rho_g \pi R^2 U_\infty^2}$$

The results of experimental measurements for spheres in a fluid indicate that the drag coefficient can be expressed as:

$$C_D = \frac{12}{\text{Re}}(1 + 0.251\text{Re}^{2/3})$$

where the multiplier of the term outside the parentheses is the drag coefficient for Stokes’ flow. The Reynolds number $\text{Re} \equiv U_\infty R/\nu_g$.

If Kn is not assumed to be zero, then the Stokes drag force on the particle also must be corrected for a slippage of gas at the particle surface. Experiments of Robert Millikan and others showed that the Stokes drag force could be corrected in a straightforward way. Using the theory of motion in a rarified gas, the mobility takes the form:

$$B = A/6\pi\mu_g R$$

Here, the numerator is called the Stokes–Cunningham factor. The coefficient A is

$$A = 1 + 1.257\text{Kn} + 0.400\text{Kn} \exp(-1.10\text{Kn}^{-1})$$

based on experiments. Thus, the mobility increases with increasing Kn , reflecting the increasing influence of a rarified gas molecular transfer regime.

2. Phoretic Forces

Particles can experience external influences induced by forces other than electrical or gravitational fields. Differences in gas temperature or vapor concentration can induce particle motion. Electromagnetic radiation also can produce movement. Such phoretic processes were observed experimentally by the late nineteenth century. For example, in his experiments on particles, Tyndall in 1870 described the clearing of dust from air surrounding hot surfaces. This clearance mechanism is associated with the thermal gradient established in the gas. Particles move in the gradient under the influence of differential molecular bombardment on their surfaces, giving rise to the thermophoretic force. This mechanism has been used in practice to design thermal precipitators for particles. Although this phenomenon was observed and identified as being proportional thermal gradients, no quantification of the

phenomenon was made until the 1920s. Einstein in 1924 discussed a theory for the phenomenon; others measured the thickness of the dust-free space in relation to other parameters. Much later, in the 1960s, the theory was refined and extended. The theoretical relation for the thermal or thermophoretic force F_t on a spherical particle is

$$F_t = -K R^2 (k_g / \bar{v}_g) \Delta T / \Delta R$$

where k_g is the thermal conductivity of the gas, $\bar{v}_g$ the mean thermal velocity of the gas molecules $\Delta T / \Delta R$ the temperature gradient at the particle surface, and K a factor depending on Kn and other particle parameters [$K \approx (32/15) \exp(-\alpha R / \alpha_g)$, where $\alpha \approx$ unity but is a function of momentum and thermal accommodation of molecules on the particle surface].

Similar expressions can be derived for other phoretic forces reflecting different effects of gas nonuniformities.

3. Heat and Mass Transfer

Particles suspended in a nonuniform gas may be subject to absorption or loss of heat or material by diffusional transport. If the particle is suspended without motion in a stagnant gas, heat or mass transfer to or from the body can be estimated from heat conduction or diffusion theory. One finds that the net rate of transfer of heat to the particle surface in a gas is

$$\dot{Q}_H = 4\pi R D_T (T_\infty - T_s)$$

where ∞ and s refer to free stream and surface conditions and D_T is the thermal diffusivity $k_g / \rho_g C_p$ (C_p is the specific heat of the particle). For mass transfer of species A through B to the sphere,

$$\dot{m} = 4\pi R D_{AB} (\rho_{A\infty} - \rho_{AS})$$

where D_{AB} is the binary molecular diffusivity for the two gases and $\rho_{A\infty}$ and ρ_{AS} are the mass concentrations of species A far from the sphere and at the sphere surface. This relation is basically that attributed in 1890 to Robert Maxwell. His equation applied to the steady-state evaporation from or condensation of vapor component A in gas B on a sphere. Maxwell's equation is analogous to Stokes' law.

The applicability of Maxwell's equation is limited in describing particle growth or depletion by mass transfer. Strictly speaking, mass transfer to a small droplet cannot be a steady process because the radius changes, causing a change in the transfer rate. However, when the difference between vapor concentration far from the droplet and at the droplet surface is small, the transport rate given by Maxwell's equation holds at any instant. That is, the diffusional transport process proceeds as a quasi-stationary process.

When the particle is moving relative to the suspending fluid, transport of heat or matter is enhanced by convective diffusional processes. Under conditions where the particle exists in a rarified medium ($Kn \gg 0$), the heat and mass transfer relations are modified to account for surface accommodation or sticking of colliding molecules and the slippage of gas around the particle.

4. Accelerated Motion

When particles are accelerated in a gas, their motion is governed by the balance between inertial, viscous, and external forces. An important characteristic scale is the time for an accelerated particle to achieve steady motion. To find this parameter, the deceleration of a particle by friction in a stationary gas is considered. In the absence of external forces, the velocity of a particle (q) traveling in the x direction is calculated by:

$$(dq/dt) + \mathcal{A}U_\infty = 0$$

or

$$q = q_0 \exp(-\mathcal{A}t)$$

if the initial velocity is q_0 and $\mathcal{A} = 9\mu_g / 2\rho_p R^2 A$. The distance traveled by the particle is, in time t ,

$$x = \int_0^t q dt' = \mathcal{A}^{-1} q_0 [1 - \exp(-\mathcal{A}t)]$$

The significance of $\mathcal{A}$ is then clear; it is a constant that is the reciprocal of the relaxation time for stopping a particle in a stagnant fluid. Similarly, one can show that $1/\mathcal{A}$ represents the time for a particle falling in a gravitational field to achieve its terminal speed. Note that the terminal speed $q_s = g\mathcal{A}^{-1}$. As $t/\mathcal{A} \rightarrow \infty$, the distance over which the particle penetrates, or the stopping distance L , is $q_0\mathcal{A}^{-1}$.

5. Curvilinear Particle Motion

When particles change their direction of movement, as for example around bluff bodies such as cylinders or bends in tubing, inertial forces tend to modify their flow paths relative to the suspending gas. Particles may depart from the path of gas molecules (streamlines) and collide with the larger body (Fig. 2). This is the principle underlying inertial particle collectors.

The trajectory of a particle moving in a gas can be estimated by integrating the equation of motion for a particle over a time period given by increments of the ratio of the radial distance traveled divided by the particle velocity, that is, r/q . Interpreting the equation of motion, of course, requires knowledge of the flow field of the suspending gas; one can assume that the particle velocity equals the fluid velocity at some distance r far from the collecting body.

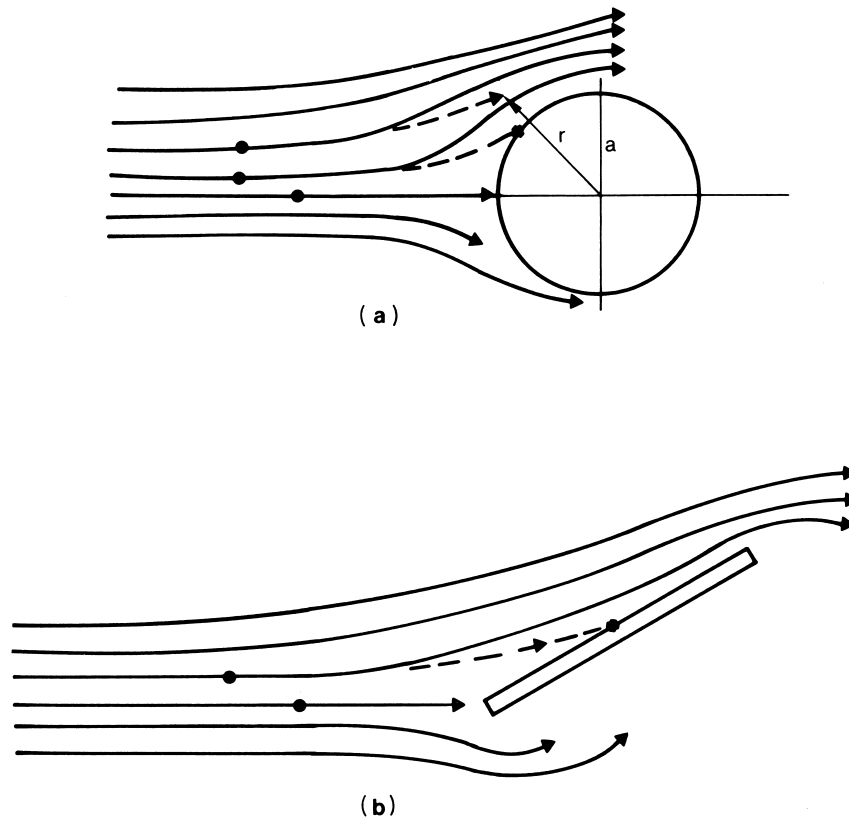


FIGURE 2 Particle motion in aerosol flow around obstacles (dashed line). (a) Flow around a cylinder of radius a ; (b) flow around a flat plate inclined at an angle to the aerosol flow.

The particle motion along curvilinear pathways and the subsequent deposition rate on nearby bodies are calculated from dimensionless particle force equations. A key parameter that derives from these equations is the Stokes number,

$$\text{Stk} \equiv \frac{U_\infty}{\mathcal{A}a} = \frac{2U_\infty \rho_p R^2}{9\mu_g a} = \frac{L}{a}, \quad \text{Kn} \rightarrow 0.$$

Stk is the ratio of the stopping distance L and a , the radius of the obstacle.

For “point” particles governed by Stokes’ law, the Stokes number is the only criterion other than geometry that determines similitude for the shape of the particle trajectories. To ensure hydrodynamic similarity, in general, the collector Reynolds number also must be preserved, as well as the ratio $I \equiv R/a$, called the interception parameter. Then, the collection efficiency of particles hitting an obstacle such as a cylinder has the form:

$$\begin{aligned} \eta &= \frac{\text{number of particles collected}}{\text{number of particles in a cross-sectional area equal to the obstacle area facing the aerosol flow}} \\ &= f(\text{Stk}, \text{Re}, I) \end{aligned}$$

Here, the interception parameter effectively accounts for a small additional increase in number of particles in a cross-sectional area equal to the obstacle area facing the gas flow, which modifies the collection efficiency to account for the finite size of the particles. The inertial collection of particles is called impaction.

6. Diffusion Processes

So far we have concentrated on the behavior of particles in translational motion. If the particles are sufficiently small, they will experience an agitation from random molecular bombardment in the gas, which will create a thermal motion analogous to the surrounding gas molecules. The agitation and migration of small colloidal particles has been known since the work of Robert Brown in the early nineteenth century. This thermal motion is likened to the diffusion of gas molecules in a nonuniform gas. The applicability of Fick’s equations for the diffusion of particles in a fluid has been accepted widely after the work of Einstein and others in the early 1900s. The rate of diffusion depends on the gradient in particle concentration and the particle diffusivity. The latter is a basic parameter directly

analogous to a molecular diffusivity. Using the theory of Brownian motion, Einstein derived the relationship for particle diffusivity:

$$D_p = kT/\mathcal{A}m_p = BkT$$

Here, k is Boltzmann's constant and m_p particle mass. In analogy to a simple kinetic theory of gases, the definition of a mean free path for particles is $\lambda_p = \bar{v}_p \mathcal{A}^{-1}$. The average thermal speed of particle is

$$\bar{v}_p = \left(\frac{8kT}{\pi m_p} \right)^{1/2} \quad \text{and} \quad D_p = \frac{\pi}{8} \bar{v}_p \lambda_p$$

Some characteristic values of these aerosol transport properties of particles in air are listed in Table II and Fig. 1.

7. Diffusion in Flowing Media

When aerosols are in a flow configuration, diffusion by Brownian motion can take place, causing deposition to surfaces, independent of inertial forces. The rate of deposition depends on the flow rate, the particle diffusivity, the gradient in particle concentration, and the geometry of the collecting obstacle. The diffusion processes are the key to the effectiveness of gas filters, as we shall see later.

A particle agitation analogous to Brownian motion is induced by turbulence in an aerosol. Turbulence is a form of eddying fluid motion often observed in atmospheric clouds or swirling cigarette smoke. The agitation caused by turbulence creates a concentration gradient and an apparent diffusion rate of particles that is much larger than that experienced in thermal motion. The characteristics of turbulent diffusion of particles is described by theory for random motion analogous to that for Brownian motion.

When particles experience a mean curvilinear motion and also have Brownian agitation, they are deposited on obstacles by both mechanisms. For very small particles of radii less than $0.1 \mu\text{m}$, Brownian motion dominates particle collection on surfaces. For larger particles, inertial forces dominate. An example of the difference in collection efficiency for spherical collectors of different size is shown in Fig. 3 for different particle diameters and aerosol flow velocity.

For surfaces oriented perpendicular to an external force, additional deposition takes place by motion induced by this force field. Examples include gravitational and electrical fields.

B. Behavior of Particle Clouds

Particle clouds are active kinetic systems. If condensable vapors are present, new particles will be formed or existing particles will grow or shrink, depending on the

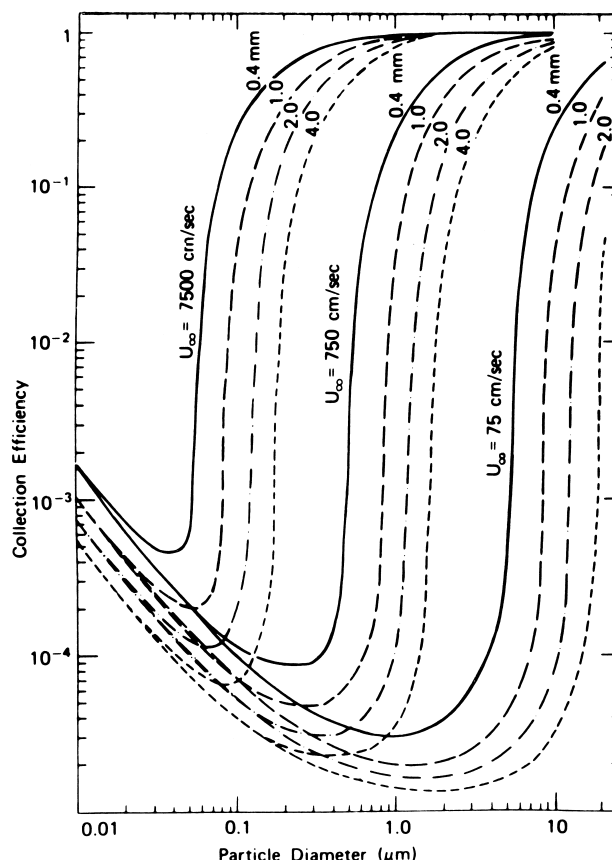


FIGURE 3 Particle collection efficiencies for spheres of different diameter and different particle diameters and aerosol flow velocity (U_∞). Particle diameters are given by the values at the upper right of each curve. The diffusion regime is at the left and the inertial impaction regime is at the right.

conditions of vapor supersaturation. Furthermore, the particles will collide with one another. The process of collision and sticking is called coagulation. These two types of processes give rise to continuously changing size distributions. The size distributions are also influenced by the deposition of particles on surfaces through diffusion, fallout, or inertial impaction. The dynamic character of aerosol clouds is crucial to their behavior and stability.

1. Nucleation and Growth

The production and growth of particles in the presence of condensable vapors is a major dynamic process. A considerable body of literature has accumulated on the subject, beginning with the thermodynamics of phase transition and continuing with the kinetic theory of molecular cluster behavior.

The process of phase changes to form clouds of particles can be induced such that supersaturation is achieved

by (1) an adiabatic expansion of a gas, (2) the mixing of a warm, moist gas with a cool gas stream, or (3) chemical reactions producing condensable species. In the absence of particles in a condensable vapor, particles are formed by homogeneous nucleation on molecular clusters in a supersaturated vapor. When vapor supersaturation takes place in a multicomponent system, mixed particles may form from heteromolecular nucleation. When particles exist in a supersaturated vapor mixture, they act as nuclei for condensation. Perhaps the best known of these processes is the formation of clouds in the earth's atmosphere by water condensation on dust or water-soluble aerosol particles.

The theory of nucleation begins with the consideration of vapor-liquid equilibrium. The vapor pressure p_0 over a flat liquid surface at equilibrium is given by the Clapeyron equation,

$$\frac{d \ln p_0}{dT} = \frac{H\mathfrak{V}}{\mathcal{R}T^2}$$

where H is the molar heat of vaporization and $\mathcal{R}$ the gas constant. At saturation, the vapor pressure p_0 is its equilibrium value at the temperature of the liquid beneath the vapor. Condensation may take place when the vapor is supersaturated or when the supersaturation ratio

$$S = p/p_0 > 1$$

The vapor pressure over a pure liquid droplet at equilibrium p_s depends on its radius of curvature. The Kelvin equation gives this relationship as:

$$\ln\left(\frac{p_s}{p_0}\right) = \frac{2\sigma\mathfrak{V}_m}{RkT}$$

Thus, a supersaturation ratio greater than unity is expected for small droplets at equilibrium with a condensable vapor. The logarithm of this ratio is proportional to the product of the particle surface free energy σ times the molecular volume of the liquid ($\mathfrak{V}_m$) and inversely proportional to the particle radius R .

The rate of formation of new particles by nucleation is given in terms of a theory for the production of clusters of molecules (embryos) in a supersaturated vapor. The production of embryos follows from considering a population of molecules and agglomerates of molecules in a homogeneous vapor. There is an energy barrier to producing large agglomerates that depends on the free energy of formation of a cluster containing a given number of molecules. As the supersaturation of a vapor is increased, the free energy of formation passes through a maximum value such that a critical level is attained. Beyond the critical value, embryos of critical size are generated that are stable and grow. The "steady" rate of production of embryos of critical size represents the expected production of new particles. The

production rate of embryos in a homogeneous supersaturated vapor is

$$I = C \exp(\Delta G^*/kT)$$

where ΔG^* is the free energy of formation of critical-sized embryos that do not evaporate and C is a rate factor. According to the theory as presented in 1935 by Becker and Döring, $\Delta G^* = -16\pi\sigma^3\mathfrak{V}_m/3(kT)^2 \ln^2 S$. The rate factor is

$$C = \frac{2p_0(n_A\mathfrak{V}_m)^{2/3}}{(2\pi m_A kT)^{1/2}} \left(\frac{\sigma\mathfrak{V}_m}{kT}\right)^{1/2}$$

where n_A is moles of condensate A and m_A is the molecular mass of condensate A.

The heteromolecular production of particles in a vapor mixture is estimated from a model similar to the homogeneous case above. However, the production rate depends on the energy of formation of mixed embryos, the composition of which depends on the thermodynamic properties of the mixture.

If particles (or ions) are already present in a supersaturated vapor, nucleation will take place preferentially on these particles at supersaturations far smaller than for the homogeneous vapor. In this case, nucleation takes place heterogeneously on the existing nuclei at a rate dependent on the free energy of a condensate cap forming on or around the nucleus. Heterogeneous nuclei always occur in the earth's atmosphere. They are crucial to the formation of water clouds and to the formation of ice particles in supercooled clouds.

2. Growth Laws

The droplet current I calculated by nucleation models represents a limit of initial new phase production. The initiation of condensed phase takes place rapidly once a critical supersaturation is achieved in a vapor. The phase change occurs in seconds or less, normally limited only by vapor diffusion to the surface. In many circumstances, we are concerned with the evolution of the particle size distribution well after the formation of new particles or the addition of new condensate to nuclei. When the growth or evaporation of particles is limited by vapor diffusion or molecular transport, the growth law is expressed in terms of vapor flux equation, given by Maxwell's theory, or

$$I(\mathfrak{V}, t) = \frac{n\partial\mathfrak{V}}{\partial t} = \frac{n4\pi D_{AB}R^2\mathfrak{V}_m(p_\infty - p_s)}{kT}$$

Other growth processes are also derived from theory. They include those associated with chemical reactions to form condensed species taking place either at the particle surface or within the particle volume. The growth by surface reactions or a vapor diffusion-limited process is

proportional to the particle surface area or radius squared for spheres. Volume reactions are controlled by particle volume or are proportional to the radius cubed for spheres.

3. Collision and Coagulation

Once particles are present in a volume of gas, they collide and agglomerate by different processes. The coagulation process leads to substantial changes in particle size distribution with time. Coagulation may be induced by any mechanism that involves a relative velocity between particles. Such processes include Brownian motion, shearing flow of fluid, turbulent motion, and differential particle motion associated with external force fields. The theory of particle collisions is quite complicated even if each of these mechanisms is isolated and treated separately.

The rate of coagulation is considered to be dominated by a binary process involving collisions between two particles. The rate is given by $bn_i n_j$, where n_i is the number of particles of i th size and b a collision parameter. For collision between i - and j -sized particles during Brownian motion, the physicist M. Smoluchowski derived the relation:

$$b = 4\pi(D_i + D_j)(R_i + R_j) \\ = \frac{2}{3} \frac{kT}{\mu_g} \left(\frac{1}{\nu_i^{1/3}} + \frac{1}{\nu_j^{1/3}} \right) (\nu_i^{1/3} + \nu_j^{1/3})$$

from Einstein's diffusion theory. This formula essentially comes from the fact that b is proportional to the sum of Brownian diffusion rate of the two particles. Analogous forms for b have been derived for other collision mechanisms.

To be derived rigorously, Smoluchowski's models must be corrected for gas slippage around the particles. This adds correction terms in Kn that accelerate the coagulation rate over the original estimates. Particles in a gas are often naturally charged. Bipolar charging increases the expected coagulation rate, while unipolar charging suppresses the rate. Fluid motion relative to the particles and the differential action of external forces induce relative motion between particles. This motion also increases the coagulation rate of particles.

IV. PRODUCTION OF AEROSOLS

A large amount of effort has gone into the investigation and development of aerosol generation devices. Over the years, a wide variety of methods for the production of aerosols has emerged; these methods depend on the technological requirements of the aerosol. For many scientific applications they include (1) control of the particle size distribution, (2) stability of operational performance for

key periods of time, and (3) control of volumetric output. The generation devices themselves have also been investigated extensively to verify the physicochemical processes in particle formation.

The generation of aerosol requires the production of a colloidal suspension in one of four ways: (1) by condensing out small particles from a supersaturated vapor (the supersaturation may come from either physical or chemical transformation); (2) by direct chemical reaction in a medium such as a flame or a plasma; (3) by disrupting or breaking up bulk material, including laser ablation; or (4) by dispersing fine powders into a gas. In each of these broad groupings, a wide variety of ingenious devices have been designed, some of which employ hybrids of two or more of these groups.

A. Nucleation and Condensation Processes

The means for the production of particles during condensation is well represented by the generator introduced by V. K. LaMer in the 1940s. This device was specifically built to produce a laboratory aerosol with controlled physical properties, using a low-volatility liquid such as glycerine or dioctyl phthalate. The device generated particles from a vapor supersaturated by mixing a warm, moist vapor with a cooler gas. Later, a wide variety of refinements of the LaMer concept emerged, including a series reported by Milton Kerker in the 1970s. Many generators using the condensation process have appeared; some of these achieve vapor supersaturation by adiabatic expansion in the vapor, others by the mixing process. Aerosols have also been formed from condensation of a supersaturated vapor produced by chemical reaction. Some examples include reactions in combustion processes, photochemical processes, and through discharges between volatile electrodes. An example of a hybrid of condensation and breakup or vaporization is an exploding metal wire technique.

Another process involves molecular aggregation by means of direct chemical reactions akin to polymerization. The best known example of this is the process of carbon particles in a premixed acetylene-oxygen flame. Evidently particle formation in this case does not involve condensation from a supersaturated vapor, but proceeds directly through the pyrolysis of the acetylene, forming in the process unstable polyacetylenes as intermediates in the flame.

Molecular aggregation to produce very small particles can be achieved through synthesis of particles in plasmas as well as ablation of bulk material using lasers. Plasma synthesis offers opportunities for high volume throughput for a wide range of refractory materials and can produce high-density particles with rapid quenching of the

plasma. Laser ablation applications can provide non- (thermodynamic) equilibrium vaporization controlling the material stoichiometry, as well as control of particle crystalline structure through temperature and concentration management.

B. Comminution Processes

The disintegration of coarse bulk material into colloids is accomplished by three main types of devices. The first is the air blast or aerodynamic atomizer, in which compressed gas is ejected at high speed into a liquid stream emerging from a nozzle. This type of breakup is found, for example, in paint spray guns, venturi atomizers, and other practical sprayers. A second class of atomizer depends on centrifugal action wherein a liquid is fed into the center of a spinning disk, cone, or top and is centrifuged to the outer edge. Provided that the flow rate of liquid into the spinning device is well controlled, sprays produced in this way are rather uniform in size, in contrast to those produced by other methods of atomization. A third type of atomizer has a hydraulic design in which a liquid is pumped through a nozzle and, upon its exit from the orifice, breaks up into droplets. Disintegration here depends largely on the physical properties of the liquid and the ejection dynamics at the nozzle orifice rather than on the intense mixing between the liquid and the surrounding gas. Perhaps most well known of these devices is the swirl chamber atomizer, which has been used in agricultural equipment, oil-fired furnaces, internal combustion engines, and gas turbines. In addition to these three main classifications, special methods are available including the electrostatic atomizer and the acoustic atomizer. The former makes use of a liquid breakup by the action of electrostatic forces, while the latter applies high-intensity sonic or ultrasonic vibrations to disrupt a liquid.

The generation of dust clouds by the dispersal of fine powders is a straightforward method, in principle. All such generators depend on blowing apart a bed of finely divided material by aerodynamic forces or by a combination of air flow and acoustic or electrostatic vibrations. The size of particulate suspensions produced in this way is limited by the minimum size of material ground up by mechanical or other means and the nature of cohesive and adhesive forces acting between particles in agglomerates. Generally, dust clouds produced by powder dispersal are not less than a few micrometers in radius.

Aerosols composed of solid particles or nonvolatile liquids with sizes much less than those attainable by atomization of pure liquids or by dispersal of powders may be produced by atomizing salt solutions. Breakup of suspensions of a volatile carrier liquid, in which solid particles, an immiscible liquid, or a nonvolatile solute are suspended,

yields very small particles after evaporation of the volatile liquid.

V. MEASUREMENT PRINCIPLES

The measurement of the physical and chemical properties of particle suspensions has been a central theme of aerosol science since its beginnings. The variety of devices and methods adopted for such purposes represents a diverse collection of instrumentation designed for specific applications. The reason for this diversity is that no single technique or group of techniques provides a means of characterizing properties over the extremely wide range of particle size, shape, and chemical composition found in nature or in the laboratory. The devices range from simple instruments for the measurement of light transmission, to porous filters for the collection of material in order to determine mass concentration, to sophisticated sensors or collectors for the characterization of particle size distribution and chemistry.

To characterize adequately the dynamic properties of a chemically reactive aerosol, a very large amount of information is required. However, aerosol properties generally are determined in only a limited way because of limitations of available techniques. With air pollution monitoring and the driving force of progress in the development of theory, heavy emphasis has been placed on the size distribution and its moments, as well as the chemical composition of particles and the suspending gas.

The measurement of particles or particle collections is achieved by one of two approaches: (1) *in situ*, or quasi-*in situ*, continuous observation or (2) collection on a medium and subsequent laboratory investigation of the accumulated particulate material. No single method provides a self-consistent, complete physical picture of a particle suspension. For example, the first approach basically assumes that the particles can be treated as inert spheres during the measurement process. The second assumes the accumulation of material on a medium or substrate without modification of the particles. This is known to be a less than satisfactory assumption for particles reacting with the suspending gas, but no better techniques have been developed.

If one focuses on the particle size distribution function as a central framework for describing aerosols, one can conveniently classify the measurement instruments according to the properties of the size distribution function. Organization of instrumentation gives perspective on the ideal requirements as contrasted with the practical limits imposed by current technology. An idealized hierarchy was suggested by [S. K. Friedlander in 1977](#). As an ideal, the modern aerosol analyzer gives a continuous

resolution in particle size with chemical composition. The ideal instrument operated at full capacity would measure and read out directly the particle distribution as a function of size and chemical properties. Only recently have analyzers employing mass spectroscopy begun to realize this potential. In practice, a variety of instruments are available that report size distribution functions or integral average properties of the distribution function. These include the single particle counter, which measures particles over discrete size ranges using differences in light-scattering properties with particle size, and devices such as the electrical mobility analyzer, which measures particles in size groups by counting charged particles in a given size range over a discrete time interval. This instrument depends on the fact that particles realize a unique equilibrium charge as a function of size. Finally, a series of devices integrates the distribution function and gives information about certain moments of the size distribution. These include: (1) idealized total particle counters such as a nuclei counter that relies on the nucleation of supersaturated vapor to produce droplets visible in a light beam, (2) a particle collector such as an impactor that segregates the sample by size over certain discrete size and time intervals through inertial forces, and (3) a total mass–chemical analyzer such as a filter placed in an aerosol stream and later submitted to the laboratory for gravimetry and chemical assay.

Once collected, particles can be sized by a variety of means. Optical and electron microscopy are probably best known and are quite reliable. Yet they involve tedious scanning of many samples to obtain sufficient counts to provide meaningful particle statistics. Microscopic techniques are suitable for solid particles and for nonvolatile liquids. Volatility creates a significant uncertainty unless the particles are trapped in a substrate that reveals a “signal” of the impacted particle.

Microscopy remains the principal standard method of particle sizing and of shape and morphological classification. Though often tedious and time consuming in its application, it remains a standard by which individual particles can be classified with confidence and most particle sizing methods are referenced.

Sampling Design

There are many pitfalls in measuring the properties of aerosols. One of the most critical is sampling of particulate matter without disturbing the aerial suspension. There are some optical devices that make measurements of an aerosol *in situ* without disturbance. However, most devices requires that a small sample be taken from the gas–particle suspension. Because of inertial forces acting on particles, it can be deduced readily that siphoning part of the fluid

as a sample must be done carefully to avoid preferential withdrawal of particles of different sizes. Particle deposition on the walls of the sampling tube, as well as possible reentrainment, must be minimized and accounted for. Care must also be taken to avoid condensation or chemical reactions in the sampling duct. Problems of this kind are especially severe in sampling high-temperature, moist gases from a stack or moist gases from a chemical reactor. Condensation can be avoided with a probe heated to the sampled gas temperature if the pressure difference has been minimized. When pressure differences in the sampler are large, control of pressure may be important. Chemical reactions on the wall of the sampling tube are often difficult to control but can be minimized using tubes lined with inert coatings such as Teflon.

The ideal condition for sampling is one in which the gas particle suspension is drawn into the instrument at a speed nearly equal to that of the external flow. Ideally, sampling should be done *isokinetically*, or with the sampler inlet velocity equal to the mainstream velocity. Only in the isokinetic case will the inertial deposition at the sampler tip be minimized and preferential size separation be small during sampling.

A. Inertial and External Forces

1. Electrical Charging and Particle Size

Two charging methods have been adopted to develop particle measurement devices. These involve diffusion charging and contact charging. Three characteristics of ion charging affect the usefulness of a diffusion charging method for aerosol sizing by electric methods. First, the relationship between electric mobility and particle size must be established. This basically provides a means of calculating the migration velocity of particles under the influence of an electrical force, which in turn gives the basis for locating a particle collection in an instrument. The mobility versus particle diameter curves are single-valued for bipolar diffusion and unipolar diffusion. This is not the case for the field charging. Second, the fraction charged must be known. Particles that do not acquire a charge during their passage through a charger cannot be influenced by subsequent electric fields and therefore cannot be measured by electrical migration. The fraction charged, in combination with aerosol losses, is the principal factor that limits the lower useful size detection of an instrument. Third, the discrete nature of electrical charge on a particle must be accounted for in the instrument output.

In principle, calculation can correct for this effect on a measured size distribution, but the methods have not been evaluated yet. If measurements are made using only the singly charged particles, then the resolution is as good as

the resolution of the mobility analyzer itself. This requires, however, that the fraction of aerosol carrying unit charge be known precisely. Aerosol concentration measurement using electrical effects requires a method of detecting the charged aerosol. This is usually done by measurement of an electrical current on a grounded collector with attachment and charge transfer of a particle. In addition, electrical sizing methods require a precipitator or classifier by which the particles of different electric mobilities are separated before detection. Various approaches have been discussed, including condenser ion counters, denuders, and ion capture devices.

2. Inertial Impaction

In the methods discussed so far, continuous observations in terms of particle size have been involved, giving detailed information on the particle concentration–size distribution but limited detail on particle morphology. An important requirement for aerosol experimentation is the ability to sample and collect particles with size segregation. One such method of sampling uses the variation of inertial impaction with mass (or size). Devices that have been designed for this purpose are called impactors. They operate on the idea that a large particle tends to collide with a surface when particle-laden air is directed to a surface, while small particles follow the gas flowing around the collector. In a typical device, the air is forced through a converging nozzle and ejected onto a plate oriented normal to the gas flow. The gas streamlines bend sharply inside, while particles with sufficient inertia hit the plate. The basic design parameters of the impactor are the nozzle throat diameter or width and the distance from the nozzle exit plane to the plate.

By operating several impactor stages at different flow conditions, one can classify the aerosol particles into several size ranges from which the size distribution is determined. These single stages can be operated in a parallel or in a series arrangement. In the parallel flow arrangement, each of the stages classifies the airborne particles at different cutoff sizes, so that the difference in the amount of the deposit on any two stages gives the quantity of particles in the particular size interval defined by the respective cutoff sizes of the two stages. In the series arrangement, also known as the cascade impactor, the aerosol stream is passed from stage to stage with continually increasing velocities and decreasing particle cutoff sizes. Thus, each stage acts as a differential size classifier. Of the two flow systems, the cascade arrangement is by far the most popular, as is evident from the large number of commercial cascade impactors currently available.

In the conventional impactor, the jet is formed in a nozzle (internal flow) and then impacts onto a plate. It is also

possible to pass the impaction plate through the particle-laden air (external flow). The effectiveness of particle collection in the latter arrangement is comparable to that of conventional impactors. In operation, these impactors normally consist of impaction plates (or cylinders) mounted at the ends of rotating arms. As the arms are rotated through the air, particles are impacted onto the collection surface. The size of the particles collected depends on the speed and width (or diameter) of the impaction surface as well as the size and density of the particles. These devices can be used to collect particles larger than 10–20 μm in diameter. Thus, for the collection of large particles, which may be difficult to sample efficiently in a conventional impactor, this type of impactor is a suitable alternative.

3. Centrifugation

The deposition of particles can be achieved by introducing external forces normal to the flow of an aerosol. This is basically the principle of size separation devices employing centrifugal forces acting on the particles. Two types of particle samplers have emerged in this group. The first are cyclones, which are passive in nature, inducing spinning air motion and forcing particles to move outward to a collection surface. The second are centrifuges in which air is spun mechanically, causing particle migration and deposition on the outer walls of the device.

Experimental investigations to determine the aerodynamic equivalent particle size for nonspherical particles led to the design of centrifugal instruments to resolve individual submicron particle deposition by the influence of an external force. The first aerosol particle size spectrometer actually providing a continuous size spectrum in terms of aerodynamic diameters was built in 1950 by Sawyer and Walton. Their centrifugal device, called a conifuge, deposited the particles according to their aerodynamic diameter in a size range between 0.5 and 30 μm on the outer wall of a rotating conical annular duct.

In reviewing the situation of centrifugal size spectrometry and after assessing the limitations of a semidispersive, cone-shaped, helical-duct, aerosol centrifuge, workers suggested that the performance of the conifuge-type size spectrometers could be improved by employing ring-slit aerosol entrances in modified designs featuring slender cones or cylindrical annular ducts. It was anticipated that ring-slit aerosol inlets would permit increased sampling rates as desired for many practical purposes. The cylindrical design would have the additional benefit of facilitating an exact theoretical performance evaluation. An actual instrument of the latter kind was subsequently built in 1968 by Berner and Reichelt. They showed that the experimental deposit patterns did, in fact, follow theoretical predictions.

In the following years, a variety of ring-slit centrifuges of the conifuge concept as well as the first spiral duct centrifuge were built and tested. A comparison of the performance tests of these devices indicated that from almost all practical viewpoints the concept of the spinning spiral duct was superior to the other designs.

The theoretical basis of the cylindrical centrifuge is a straightforward application of force balance on particles in the annulus. If the centrifugal force acting on the particle is constant, the length from the entrance where a particle of given size is deposited is proportional to the aerosol flow rate, but inversely proportional to rotation speed and the square of the particle radius. These relationships are borne out by deposition experiments using particles of known radius and density.

4. Diffusion and Filtration

Collection on porous filter media is perhaps the most efficient means of particle removal. Aerosol filtration is an effective means of air purification, while at the same time it has been widely used for sampling airborne material for mass and chemical composition determination. A wide variety of filter media is available, ranging from fibrous mats of relatively inert material to porous membranes. Fibrous mats and model filter arrays appear microscopically as stacks of overlaid cylinders, where the cylinders may be smooth or rough. In contrast, the membrane media are plastic films with microscopic holes of nearly uniform size; nuclepore filters, for example, are produced of sheets of polyester, and the holes are introduced by neutron bombardment.

Fibrous filters are the most economical and effective devices for the purification of air from suspended particles. This purification is achieved with minimal loss of pumping energy associated with flow resistance, compared with other types of filters. The porosity of such materials is 85–99%, and fiber diameter varies from 10^2 to 10^{-2} μm .

The advantage of membrane filters is that particles do not become imbedded in the filter medium. Thus, individual particles are readily identifiable and characterized microscopically on the filter surfaces. Furthermore, certain kinds of chemical analysis, such as X-ray fluorescence analysis, readily can be done *in situ* with minimal effects of filter interference on the membrane substrates.

Sampling devices range from simple filter holders to sequential configurations for automated routine air monitoring of many samples in series. Membrane filters can be obtained in different pore sizes, so that they can be used in series as particle size fractionators.

The theory of filtration is a direct application of principles of Brownian diffusion discussed previously. The objective of the theory is to provide a framework for cal-

culating the number of particles of a given size deposited per unit area or unit filter length, as the sample depends on flow rate, porosity per hole or filter diameter, temperature, pressure, presence of condensation or chemically reactive vapors, electrical fields, and so on. The overall filter collection efficiency, combined with the pressure drop or flow resistance, is a crucial characterization parameter for the selection of appropriate filters for air purification.

Particles are deposited from a gas layer adjacent to the substrate. Deposition takes place by convective diffusion and interception. Thus, the complex pattern of flow through a filter becomes a key to calculating its efficiency. In principle, one calculates the flow through a fibrous filter in terms of a superposition of flow around a cylindrical array, taking into account the mutual interactions between fibers using the packing density.

The character of flow through a fibrous mat can be seen by examining the drag force on a unit fiber length in terms of pressure drop across the filter.

The superposition of electrostatic forces on particle behavior near a filter mat can have appreciable influence on filtration efficiency. The deposition patterns can take on significant treeing or branching of agglomerates on individual fibers. This aerodynamically distorts the cylindrical collector surface and branches the surface area, as well as distorting the electrical field around the collector.

For air-monitoring purposes, gravimetric measures of total mass concentration from filters, combined with chemical assessment, generally require a relatively large amount of sample. Also, as will be seen later, separate samples free of influence of chemical interactions during collection are of interest. A device for monitoring applications was developed in the 1970s that improves on the high-volume sampler. The device is called dichotomous sampler. It collects particulate material in two size groups, between 2 and 5 μm diameter and less than 2 μm diameter. Segregation of very large particles (>10 μm) is readily achieved by design of an inlet shroud, which restricts entry of particles larger than 10 μm diameter. Separation of the coarse and finely divided particles is achieved by a method called virtual impaction. In principle, this method avoids such difficulties as particle bounce-off, re-entrainment, and nonuniform deposition. In addition, it provides a separation of large and small particles such that they cannot chemically interact with one another after collection on a substrate. This sophistication in sampling is important for characterizing chemically unstable particles in air.

Virtual impaction uses the principle of inertial separation, but the impaction plate is replaced by a zone of relatively stagnant air below the nozzle. The virtual surface formed by deflecting streamlines gives separation conditions similar to those in conventional impactors. Large particles travel straight through into the low-flow region,

while the small particles bend with the high-speed flow as it moves radially around the receiving tube. Fractions of different sizes are then deposited on separate filters.

Instead of using the virtual impactor approach, North American air monitoring programs in the 1980s and later have adopted “simpler” reference methods that use the weighing of filters in the laboratory. The filters are obtained from samplers equipped with an inlet device that provides for a sharp cut-point in particle entry for samples of particles $<10\ \mu\text{m}$ diameter or $<2.5\ \mu\text{m}$ diameter, which are operated over a fixed time period of 24 hours. The inlet fractionation is facilitated either by a carefully designed cyclone or by an impactor. The combination of the two samplers can give estimates of mass concentration for fine-particle and coarse-particle concentrations.

Recent advances in continuous mass monitors may replace the labor-intensive filter methods for air monitoring. One of the promising devices for continuous monitoring is the tapered element oscillating filter measurement (TEOM). This method measures the change in natural oscillation frequency of a suspended filter, which ideally is proportional to the mass of particles collected. TEOM instruments have been deployed in the 1990s at selected sites in North America and are undergoing extensive intercomparison with the gravimetric filter method.

5. Light Extinction and Optical Devices

Small particles scatter and absorb light. This phenomenon has been used to investigate aerosol behavior extensively since Tyndall’s work in the nineteenth century. In more recent years, instruments have been built to take advantage of light interactions to deduce particle size distributions. To appreciate how such devices work, we introduce certain basic principles of light interaction with airborne material.

Basically, the scattering and absorption of light by individual particles depend on their size and shape, their index of refraction, as well as the wavelength of incident light. The total scattering and absorption from a beam of light by an aerosol cloud corresponds to the summation of the scattering from all particles of different size and refractive index. If the particle cloud is dilute enough, the effects of multiple light scattering can be disregarded, and the summation of single particles suffices to describe the interaction.

The light attenuation process can be analyzed by considering a single particle of arbitrary size and shape irradiated by a planar electromagnetic wave. The effect of the presence of the particle is to diminish the amplitude of the plane energy wave. At a distance large compared with the particle diameter and the wavelength, the scattered energy appears as a spherical wave centered on the particle and possessing a phase different from the incident beam. The total energy lost by the plane wave, the

extinction energy, is equal to the scattered energy in the spherical wave plus the energy of absorption.

The light intensity is the conventional measure of the energy of scattered light and, in cgs units, has units of $\text{erg}/\text{cm}^2\ \text{sec}$. The intensity of scattered light is proportional to the intensity of the incident light beam I_0 and the radial distance expressed as:

$$I = I_0 F(\theta, \phi, \lambda) / x^2$$

where θ, ϕ are angular coordinates, λ the wavelength of light, and $x = 2\pi r / \lambda$. In general, $F(\theta, \phi, \lambda)$ depends on the wavelength of the incident beam and on the size, shape, and optical properties of the particle but not on r , the radial distance from the particle. For spherical particles, there is no dependence.

The scattering function can be determined from theory for certain important special cases. The performance of optical single-particle counters depends on the variation of the scattering function with angular position.

Rayleigh scattering for $x \ll 1$ and the large-particle extinction law for $x \gg 1$ provide useful limiting relationships for the efficiency factor. Aerosol light scattering, however, is often limited by particles whose size is of the same order as the wavelength of light in the optical range from 0.1 – $1\ \mu\text{m}$ in diameter. In this range, Rayleigh’s theory is not applicable since different parts of the particle interact with different portions of an incident wave. Yet, such particles are still too small for the large-particle scattering theory to be applicable. In such a situation the theory of Mie is applied. Expressions for the scattering and extinction are obtained by solving Maxwell’s electromagnetic theory for the regions inside and outside a homogeneous sphere with suitable boundary conditions. Mie found that the efficiency factors are functions of x and the index of refraction alone. The calculations must be carried out numerically, and the results have been tabulated for specific values of the refractive index.

The intensity function in itself is not sufficient to characterize the scattered light. Also needed are the polarization and phase of the scattered light. For measurement applications including instrument design, the parameters of most interest are the intensity function and the scattering efficiency.

6. Single-Particle Optical Analyzers

Particle sizing by means of light scattering on single particles was understood by 1900. In the meantime, the subject has been steadily developed. Since about 1960, optical particle counters using white light illumination have been commercially available. After the invention of the laser principle, several attempts were made to replace the white light illumination of scattering devices by coherent and monochromatic laser light illumination.

To examine the influence of the optical properties and the shape of a particle on the response of a light-scattering device, it is helpful to introduce some definitions of equivalency in particle size. Particle-scattering cross section is related to the geometric radius or diameter by some measured relation, which depends on a calibration refraction index and spherical shape of the particles. For a nonspherical particle, a light-scattering diameter is defined that is related to equivalent spheres, but the particle shape is described by an optical shape factor. Thus, an optical counter measures a light-extinction cross section and not particle size directly. Unless the counter is calibrated for a given aerosol particle material, its optical size generally will not be the same as the geometric size.

The size resolution ideally depends on the monotonic nature of the response curve. Since the counter response is strongly dependent on the index of refraction of particles of a given size, the size discrimination will be quite variable for an aerosol of heterogeneous composition.

Single-particle optical analyzers are especially useful for continuous measurement of particles of uniform physical properties. However, as discussed earlier, uncertainties develop in the measurement of particle clouds that are heterogeneous in composition because the refractive index may vary from particle to particle. Thus, in making atmospheric aerosol measurements, workers have assumed an “average” refractive index characteristic of the mixture to estimate a calibration curve or have reported data in terms of the equivalent particle diameter for a standard aerosol, such as suspended polystyrene latex spheres.

7. Light Transmission and Nephelometry

The extinction of a light beam or incident radiation associated with a cloud of particles basically involves a measure of a moment of the particle number–size distribution roughly proportional to the surface concentration. The extinction of a light beam is well-developed basis for semi-quantitative measurement of particles suspended in gases. Devices for these purposes, including smoke photometers, have been available commercially for many years. They may take many forms, one of which is the transmissometer, or opacity instrument.

Transmissometers in simplest form consist of a light source and a detector located some distance along the axis of the light beam. Among their applications, these devices are commonly used at airports to provide data in visual ranging conditions and are used to measure particle loadings in smokestacks. When calibrated for the type of particles present in this aerosol, a semi-quantitative measure of the particulate emissions can be made, with knowledge of the volumetric gas flow.

The attenuation of a light beam is given in terms of the extinction coefficient, sometimes called the attenuation coefficient or turbidity, and it is a key measure of the optical behavior of particulate systems. In terms of the separate contributions for particle scattering and absorption,

$$b_{\text{ext}} = b_{\text{sp}} + b_{\text{ap}}$$

if the contribution of gas absorption and scattering is disregarded. Here, the particle scattering coefficient b_{sp} and the absorption coefficient b_{ap} are functions of wavelength of incident radiation. The particle light-scattering coefficient has been measured by a variety of instruments. One simple device that was invented many years ago is the integrating nephelometer. The absorption coefficient is difficult to determine but has been obtained through analysis of transmissometer data or inferred from absorptivity measured from filter-collected material.

8. Remote Sensing Techniques

The development of the relationships between scattered light and aerosols has stimulated the use of radiation transfer theory for remote sensing of particles in planetary atmospheres. Highly sophisticated experimental and theoretical techniques have emerged for the interpretation of observations of sunlight and artificial light sources in the earth's atmosphere. A description of their application depends on further development of the concepts of radiant energy transfer.

According to one simple concept, remote sensing depends on light attenuation as a function of light path length. The reduction in the intensity of the light beam passing through an aerosol is obtained by integrating between any two points along the beam, L_1 and L_2 ,

$$I_2 = I_1 e^{-\tau}$$

where the optical thickness $\tau = \int b_{\text{ext}} dz$ is a dimensionless quantity (z is the path length of light); b_{ext} has been kept inside the integral sign to show that it can vary in space with the aerosol particle concentration.

For aerosols in the atmosphere, the light-scattering coefficient empirically is found to be proportional to particle mass concentration in the range of diameter below $10 \mu\text{m}$. This relationship is useful for coarse estimates of airborne particle concentrations.

Remote sensing is achieved by the three groups of techniques listed in Table III: (1) ground-based passive optical sensors, (2) airborne passive sensors (aircraft, balloon, or artificial satellite), and (3) active sensor systems, employing a controlled or manmade light source. The instrumentation designed and built for such purposes is diverse and ingenious. It has occupied the thoughts of astronomers,

TABLE III Remote Sensing Methods for Atmospheric Aerosol Characterization

Method	Measurable light	Light source
Ground-based, passive		
Photometry and radiometry	Optical thickness; sky brightness	Sun
Polarimetry	Polarization of skylight	Sun; diffuse sky
Polar nephelometry	Extinction coefficient	Sun; skylight
Teleradiometry	Horizon brightness; relative contrast	Reflected and scattered light
Airborne, passive		
Spectrophotometry	Albedo; optical thickness	Reflected sunlight
Limb occultation (satellite) radiometry	Optical thickness; polarization	Sunlight
Active sensing		
Transmission/backscatter	Optical thickness	Searchlight; LIDAR ^a

^a LIDAR denotes light detection and ranging.

spectroscopists, meteorologists, and space engineers for many years. The publications listed in the bibliography contain the details of the methodology and the intercomparisons between remote sensors of different kinds and direct aerosol observations.

In general, the methods are difficult to interpret quantitatively in terms of aerosol properties because of ambiguities in the size distribution–concentration–distance profiles and variations in chemical properties contributing to the index of refraction. Nevertheless, remote sensing continues to be important for the surveillance of aerosol behavior in planetary atmospheres.

B. Methods for Chemical Characterization

After the collection of particles, it is useful to determine the chemical characteristics of the material. This can be accomplished in terms of analysis of a whole sample corresponding to the total mass concentration, or it can be done on a size-fractionated basis. In some cases, individual particles can also be examined. Chemical characterization is very important when one is considering a heterogeneous collection of aerosol particles such as those found in the ambient air or in the workplace. These include whole sample microscopic analysis by collected batch, as well as continuous measurement.

1. Macroscopic Techniques

Macroscopic methods for chemical analysis essentially take either all of the particulate matter sampled or a significant portion of it for bulk analysis. Traditionally, this has been approached by the application of standard microchemical techniques of wet chemistry. The unique analytical requirement for aerosol particle samples is the microgram quantities collected. The analytical methods adopted must be capable of detecting these quantities in

a matrix of many different, often unknown components. New methods have been introduced that involve spectroscopic examination by a variety of techniques including X-ray fluorescence, plasma emission spectroscopy, neutron activation analysis, photoelectron spectroscopy, and mass spectroscopy.

The wet chemical methods are summarized in Table IV. Basically this approach centers on the water-soluble extract obtained from filter substrates, impactor plates, or other collection surfaces. The extraction process has to be done with some care to ensure that all of the water-soluble material is removed. Standard extraction methods now involve the use of ultrasonic devices to maximize extraction efficiency. Once the extract is obtained, it can be subjected to a number of the methods listed in Table IV, such that a detailed elemental breakdown by inorganic and (water-soluble) organic carbon is accomplished.

TABLE IV Wet Chemical Methods for Particle Analysis

Particle	Method
Cations	
Ammonium	Indol phenol blue colorimetry, ion chromatography
Anions	
Sulfate	Methylthymol blue colorimetry, ion chromatography
Nitrate	Cadmium reduction, ion chromatography
Chloride	Ferric thiocyanate, ion chromatography
Elements	
Pb, Fe, Na, K, Ca, Cr, Ni, As, Mn, Si	Atomic absorption spectroscopy, plasma emission spectroscopy
Carbonaceous material	
Solvent-soluble organics	Extraction and carbon detection; differential thermal analysis

The methods listed yield the concentrations either of water-soluble ions measured in terms of certain oxidized states or of elements. For example, materials appearing as sulfate and nitrate may include lower oxidation states, but the methods basically do not distinguish among them. The metal elements found are either oxides or soluble salts. The actual composition of the material is indeterminant, but workers have deduced the composition suspected to be present by a material balance, combined with knowledge of the origins of the particulate material.

2. Microscopic Techniques

The hope of chemical characterization of individual particles or even the surface nature of individual particles has led investigators to apply new microtechniques to aerosol research early in their development. Early methods employed the electron microscope for the identification of particles, which were captured on a reactive substrate. The collection surface was selected to provide an indicator of a specific compound. Frank and Lodge in 1967 found that sulfuric acid particles could be identified from a satellite ring structure after collection on a silicon surface. Later, Bigg, in 1974, used specific chemical surface coatings to bring about colored chemical reactions, which could be identified microscopically. These methods are generally semiquantitative in terms of mass concentration but give a number density estimate and a rough particle size estimate, especially if they are used for qualitative identification for some years.

Electron microscopy has been used for some years. The scanning electron microscope has been used in conjunction with energy-dispersed X-ray (EDX) for the analysis of single particles. Microprobe techniques include: (1) AEM (auger electron microscopy), (2) ESCA (photoelectron spectroscopy), (3) SIMS (secondary ion microscopy), (4) EPMA (electron probe micro analysis), (5) MOLE (laser Raman microprobe analysis), and (f) LAMMA (laser microprobe mass analysis). AEM, ESCA, and SIMS are surface-sensitive methods that provide knowledge of surface properties but are difficult to extrapolate to heterogeneous materials. EPMA gives information related to bulk properties with a greater penetration depth; MOLE may also be useful in this respect. LAMMA has only recently become available as a research device but will undoubtedly be used more extensively in fine-particle characterization.

In general, microprobe analysis methods have been considered semiquantitative, but EPMA has been promising for quantitative studies. With further improvement and investigation, these microprobe techniques will be useful for the characterization of surface and bulk particle properties.

3. Continuous Methods

Continuous air monitoring for trace contaminants in ambient air has developed extensively since the mid-1960s as a result of stimulation from new air pollution measurement requirements. Workers expect that similar needs will develop as certain chemical constituents of particulate material are identified as factors in human health effects. Techniques for the continuous chemical characterization of particulate matter are slow in coming because the amounts of material sampled are small, often below the detection limits of instrumentation. In all cases so far, either a precollection method like filtration or a special detector of high sensitivity has been required.

Flame photometry has promise for the measurement of sodium, lead, and potassium. An application to measurement of sodium and alkali metals has been reported. The continuous measurement of sulfur-containing particles has received considerable attention. The motivation for observation of sulfur-containing particles comes from concern about the potential hazard posed by sulfate in the atmosphere.

Practical methods also have been reported for semicontinuous measurement of nitrate and carbon in particles from ambient air. For example, an instrument for nitrate monitoring uses collection of particles on an impactor surface, followed by flash volatilization and determination of the nitrate present using a chemiluminescence technique. Ion chromatographs also have been adopted for semicontinuous determination of gaseous and particulate nitrate. Real-time carbon analyzers also are available, one of which uses differential thermal analysis of impactor-collected material.

An important advance in continuous analyzers uses both particle size data and single-particle chemical composition. These instruments employ a method of rapid depressurization of the aerosol that produces a particle beam and irradiation of particles to generate ions that are analyzed by mass spectroscopy. The single particle analyzers have been employed in atmospheric research recently but have not reached the stage where they are used routinely in air monitoring.

VI. INDUSTRIAL GAS CLEANING

The environmental control of particle suspensions from industrial practice is an important aspect of aerosol technology. Control of industrial aerosols is done both in the workplace to preserve safe conditions and at the exit exhaust stack to minimize the pollution of ambient air. In general, a variety of regulations govern industrial operations.

These are outlined in the form of emission limits or air-borne particle concentration limits.

When the emission limitations for a planned operation are estimated to be in conformance with air quality regulations, a series of process design interactions may take place to ensure compliance. To reduce emissions of existing and new facilities, considerable innovation may be required in process and system engineering. In general, gas cleaning and atmospheric emissions can be minimized by reducing to the greatest degree possible the amount of contaminants entering exhaust gas streams. Process, operational, and system control will concentrate contaminants in the smallest possible air volumes. As an engineering principle, this is important since the cost of control equipment is based mainly on the volume of gas that has to be handled and not on the amount of material to be removed. Also, most cleaning equipment is more efficient at high concentration, all else being equal. Emission control with modern, highly efficient industrial cleaning devices can cost over \$2 per cubic meter per second of installed capacity and entails high, continuous operational and maintenance expense. Thus, innovation in the process or system control steps represents major opportunities for minimizing air cleaning control costs. Such practice can also be vital to improving employee safety inside the plant. Emission reductions or their elimination can be achieved in a variety of ways, including substituting products, changing processes, revising plant layout, or revising internal ventilation systems.

A. Particle Removal Technology

Methods for the removal of particles from industrial streams rely on the same generic collection techniques used for sampling and collecting particles for examination and characterization. These methods include: (1) gravitational settling, (2) centrifugal separation, (3) inertial capture by wet scrubbing, (4) filtration, and (5) electrostatic precipitation (see Fig. 1).

In all cases, the efficiency of removal using a control device is given by the collection efficiency, equal to the ratio amount collected in the device to the amount in the inlet gas stream. Another common measure of efficiency is the penetration, which is 1 minus the collection efficiency. The decontamination factor is the reciprocal value of the penetration. Modern gascleaning machinery aims for very high collection efficiencies exceeding 99%, at least over a given particle size range.

1. Gravitational Settling

Removal of particles by gravitational fallout is the least efficient of available techniques. It is primarily used for very large particles when the settling rate is rapid. Gravity removal is inexpensive and generally involves flowing

gas through a large plenum chamber, sometimes equipped with horizontal plates to facilitate deposition. Removal efficiency depends on the number of shelves, the size of the chamber, and the air flow rate.

2. Centrifugal Separation

Centrifugal separation or cyclones rely on inertial forces in curvilinear flow to induce deposition on the walls of the collector. The collection efficiency of such devices depends on the air flow at the inlet, the spinning rate, the collector dimensions, and the particle size. Efficiencies of 90% or greater can be achieved for particles of 10- μm diameter and larger. Cyclones have been designed with various configurations to achieve high gas rotation rates for given inlet velocities. Usually rotation is induced in a helical path such that layers of collected particles can slide by gravity down the walls to an accumulator.

3. Wet Scrubbing

Inertial collection can also be achieved effectively by taking advantage of liquid gas contactors through droplet-particle collision and contact with liquid sheets during gas passage through a liquid spray. Sprays are added to cyclones, for example, to improve the collection efficiency of the latter for small particles. Wet scrubbing also uses interception and diffusion processes as a supplement to inertial effects. Wet scrubber collection efficiencies depend on the spray droplet diameter, the particle diameter, and the flow rates of the aerosol and the liquid spray. Efficiencies of 99% or greater are achieved in scrubbers for particles of 1- μm diameter or larger.

A variety of designs have been developed to maximize the relative motion between particles and spray droplets or the contact between liquid and particles. These include jet impingement devices, packed or sieved plate towers, preformed and gas atomized spray towers, and venturi sprayers.

4. Filtration

Industrial filtration has become increasingly attractive for gas cleaning with new high-volume flow configurations; it is possible to achieve very high collection efficiencies for finely divided particles less than 1 μm in diameter. Industrial filters have efficiencies of well over 99% for particles over a full range of size. The primary design factors are pressure drop (air pumping expense) and ability to clean filters.

Filter units are broadly classified into two types: (1) fabric or cloth bag systems and (2) deep columns packed with rock or other contactors such as fiber plates or glass beads or rings. Commercial fiber materials are selected for

strength per unit mass, high temperature stability, chemical uncertainties, and cost. Fabrics include cotton, wool, paper, nylon, glass, and asbestos.

The application of external electrical fields can enhance filtration efficiency beyond a simple system. Bipolar electrostatic charge between the fabric and the particles can induce migration to the filter surface and particle agglomeration in the aerosol.

5. Electrostatic Precipitation

A particle removal method commonly used in industry is electrostatic precipitation. Industrial interest in this very efficient scheme can be traced back to 1911 with the investigation of F. Cottrell. His pioneering studies of sulfuric acid mist removal from copper smelter effluents led to the production of the Cottrell precipitator.

Success in the nonferrous metals industry was followed by the application of precipitators to the collection of dust from cement kilns. From these beginnings, the use of precipitators has expanded to include a wide variety of forms including unique boilers for electric power generation. The principal uses of precipitators today are in gas-cleaning applications in which high collection efficiencies of small particles are required for processes that emit large gas volumes. Since the separation force in a precipitator is applied to the particle itself, the energy required for gas cleaning is less than that for equipment in which energy is applied to the entire gas stream. This unique characteristic of precipitators results in lower gas pressure drops and usually lower operating costs than other methods of gas cleaning.

The precipitation process requires: (1) a method of providing an electrical charge on a particle, (2) a means of establishing and maintaining an electrical field, and (3) a method of removing the particle from the precipitator.

The process of electrically charging a particle involves the addition of electrons to or removal of electrons from the material or the attachment of ionized gas molecules to the particle. Almost all small particles in nature acquire some charge as a result of naturally occurring radiation,

triboelectric effects due to transport through a duct, flame ionization, or other processes. These charges are generally too small to provide effective precipitation, and in all industrial precipitators charging is accomplished by the attachment of electrical charges produced by an electrical corona. A corona discharge producing negative ions is normally used in precipitators.

Ideally, electrical precipitators generally achieve collection efficiencies of more than 99% for a full range of particle size. The efficiency depends on the ratio of the collector surface area particle size and dielectric properties and the volumetric gas flow rate times the charged particle migration speed induced by the applied electrical field.

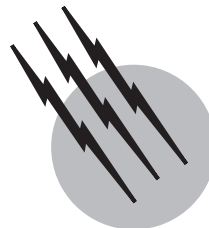
The removal of electrostatic precipitators can be improved by combining electrical collection with filters either with conventional design in series or via integration of the two technologies. Hybrid systems of this type are being introduced in some industries where very high efficiencies are needed.

SEE ALSO THE FOLLOWING ARTICLES

• ATMOSPHERIC DIFFUSION MODELING • CLOUD PHYSICS
• COAL STRUCTURE AND REACTIVITY • COMBUSTION
• FIRE DYNAMICS • LIQUID ROCKET PROPELLANTS •
PARTICLE SIZE ANALYSIS • PLANETARY ATMOSPHERES
• POLLUTION, AIR • POLLUTION, CONTROL • VOLCANOLOGY

BIBLIOGRAPHY

- Friedlander, S. K. (1977). "Smoke, Dust and Haze," Wiley-Interscience, New York.
- Hidy, G. M. (1984). "Aerosols: Industrial and Environmental Science," Academic Press, New York.
- Reist, P. (1984). "Introduction to Aerosol Science," Macmillan, New York.
- Seinfeld, J., and Pandis, S. (1998). "Atmospheric Chemistry and Physics: From Air Pollution to Climate Change," Wiley-Interscience, New York.



Batch Processing

Narses Barona

Ethyl Corporation

- I. Design and Operation
- II. Discontinuous Processing
- III. Batch Processing
- IV. Batch Processing Plants
- V. Economics of Batch Operations
- VI. Optimization
- VII. Simulation
- VIII. Analysis
- IX. Safety
- X. Materials

GLOSSARY

Batch cycle Period of time ruling the manufacturing operation of a single product.

Block cycle Period of time ruling the manufacture of the various products made in a multiproduct plant; the sum of the batch cycles of the various products.

Break-even Point of the revenue versus production level chart where the total product cost equals the total income.

Fatigue Manifestation of a cumulative process leading to progressive fracture of a material subjected to cyclic loading.

Integrator Algorithm used for the numerical solution of the problem $\int_a^b Y dx$ between the limits a and b .

Jacobian Matrix of the partial derivatives of the rates of the model variables with respect to the variables.

Jacobian banded The Jacobian is banded when the ma-

trix elements lie in bands parallel to the diagonal and the other matrix elements are null.

Modeling Procedure underlying the design and use of models to predict the characteristics of a system.

Optimization Design methodology by which the cost of making a chemical is minimized by reducing capital and operating costs.

Reactor Pressure vessel for heating and mixing chemicals and changing their chemical composition.

Scaleup Increase in size of a chemical unit (for example, a reactor), expecting the large unit to behave like the small unit.

Scheduling Setting up a production plan for the manufacture of one or more chemicals.

Simulation Development and use of computer models for the study of dynamic systems.

Stiff Refers to a problem defined by a set of ordinary differential equations. The Jacobian of the set contains

an element the magnitude of which is much larger than that of the other elements.

Train Section of a plant, or group of pieces of equipment, which operates relatively independently and in the same time length.

BATCH PROCESSING is a chemical engineering methodology for producing chemicals in cycles. In engineering practice, the term often refers to the analysis of the methodology. Industrial plants are run batchwise or continuously. There are economical or practical advantages to each method. The large size of many chemical and petroleum industries precludes the use of batch processing. The petrochemical, pharmaceutical, and food industries are batch when production levels make batch processing economical. Batch processing is suitable for processes involving multiple grades, variable market demands, or products with a short profitable life, unusual specifications, or requiring subjective tests, such as taste in the food industry; for products made at low or moderate rates, such as pharmaceuticals, fine organic chemicals, or specialty chemicals; when annual requirements can be manufactured in a few batches; or when production can be scheduled together with other products in an existing or in a new plant.

I. DESIGN AND OPERATION

Capital and operating costs are optimized by minimizing the capital investment of modular sections of a plant, by reducing the operation cycle of one product, or by making several products in the same plant to use idle equipment. The process may be operated for the whole year or during an optimum period, having idle equipment thereafter. The use of idle equipment may be resolved with multiple production, which is accomplished in multiproduct plants. The production of specialty chemicals has led to the use of the same plant for the manufacture of several products. The idea is an extension of long practice at pharmaceutical installations. Simultaneous production with the purpose of sharing equipment and reducing capital costs has led to the establishment of multipurpose plants. Certain aspects of design and operation become critical, due to the production in cycles. The reactants, products, and equipment are subjected to drastic changes in operating conditions due to the working routine of the plant. These changes may produce significant pressure or thermal stresses or situations promoting localized corrosion; the cyclic operation may cause fatigue of the materials. The combined result may lead to materials failure. These problems are not current in continuous plants, where the operating conditions are

selected to maintain smooth response of the materials and the equipment. The chances for thermal decomposition of the reagents, products, or certain intermediates and other safety risks increase significantly, since the startup conditions, where safety is more critical, are produced every time the operation cycle of the plant is repeated. Thus, in batch processing, safety is of greater concern than in continuous processing.

One of the challenges in the development and construction of new processes is the scaleup of chemical reactors. Initially, these reactors are batch, since most new processes are conceived in the laboratory. Development of continuous processes has to face the conversion from batch to continuous, and the impressive growth from the test tube and the pilot plant to industrial sizes hundreds of times larger than the experimental size. Hence, the scaleup to continuous commercial units is always full of surprises and uncertainties. In the continuous plant, the chemicals often behave differently than in the batch reactor. Secondary reactions, which are unseen in the pilot plant, become important in commercial scale, causing the undesired effect of reducing product quality and specifications. These and other challenges, which are inherent in the development of continuous plants, forced designers to return to the construction and operation of batch processes. Many uncertainties are eliminated using batch reactors, since a lack of knowledge can be covered by adjusting operating conditions until satisfactory results are obtained.

II. DISCONTINUOUS PROCESSING

The concept of batch production is as old as the history of humanity. The manufacture of soap, paints, fermented beverages, drugs, dyes, and metals are examples. The concept of large-scale manufacture, which developed after the industrial revolution, produced a change to continuous production. The goal achieved was making larger quantities of goods at a lower price. The large sizes of chemical, petroleum, and metallurgical plants could never have been possible otherwise. Interestingly, batch processing was reborn in the same facilities of the large continuous plants. Batches of raw materials and products had to be stored to meet the trends of the market demand and supply. Optimization of batch storage acquired importance for the cost of the equipment and the value of the materials held. Continuity was often disrupted by variability of the quality of the raw materials due to the need for processing stocks from different origins in the same plant. The use of catalysts became another cause of discontinuity, since the catalytic beds had to be dumped or regenerated periodically. Now discontinuous processing is used when continuous production is not economical because the reactions and

other processing steps are too slow or the manufacturing sequence is too complex. Many chemicals, petrochemicals, drugs, and vitamins are manufactured batchwise. The unit operations of continuous processing are practiced batchwise in batch processing; descriptions of these can be found in standard references. Some of the advantages of continuous processing are retained. For example, solid materials can be charged on a batch basis, while the fluid reactants are added continuously, thus minimizing conditions for side reactions, or the feed rate can be scheduled to favor the kinetics of the main reaction and increase product yield. In modern batch plants manual operations are refined by computer-controlled sequencing, and specialized criteria for product finishing are applied from batch to batch. Thus, product quality is more uniform and not as greatly affected by the differences in the skills of plant operators. Three types of discontinuous processes are common: batch, when materials do not enter or leave the unit during the cycle; semibatch, when one or more but not all components enter or exit the unit during the cycle; and semicontinuous, characterized by a processing rate at which the unit runs continuously, subject to periodic startups or shutdowns.

III. BATCH PROCESSING

Batch processing is a form of discontinuous processing characterized by operation in uniform, repetitive cycles. The batch cycle time is the time elapsed between successive batches (Table I). The reaction time is only a fraction of the total cycle time; additional time must be provided for charging the raw materials, bringing them to reaction conditions, allowing the reaction to take place, treating the reaction mass once the desired conversion is attained, discharging the products, cleaning the reactor when it is

needed, time losses due to human error or inefficiency, and other operations that may be required. Table I illustrates two batch polymerization cycles where the reaction time is almost 60% of the batch cycle time.

A plant operability analysis helps to establish the correct cycle time. In certain situations, the batch cycle of the entire plant is not determined by the units around the main reactor. Additional time may be needed downstream for product finishing in equipment that operates in series with the main reactor. The batch cycle time T is a composite of three contributions: (1) the sum of the batch residence times $t(i)$ in the M true batch units of the plant that operate in series; (2) the sum of the residence times $t(j)$ in the N semicontinuous trains of the plant that operate in series; and (3) the sum of the downtimes $t(k)$ in series encountered in the total batch cycle:

$$T = \text{SUM } t(i) + \text{SUM } t(j) + \text{SUM } t(k) \quad (1)$$

IV. BATCH PROCESSING PLANTS

A plant is a battery of equipment for the conversion of raw materials to products. It consists of three main sections: storage, process equipment, and utilities. Process equipment for batch processing is standardized so it can be adapted to make different products. The major processing units of a batch plant are operated on a batch or semibatch basis. Transfer interunits (pumps) run on a semicontinuous basis. Heat and mass-transfer operations are continuous, but in some instances they are semicontinuous or batch. Storage is provided to keep the raw materials, to hold the product and rework streams, and to serve as a buffer between the batch and the continuous sections of the plant. It provides for surge vessels, which decouple the upstream and downstream units and maintain continuity of operation. It may require a substantial portion of the plant capital investment. Utilities are the section that supplies the plant with water, fluids (air or nitrogen) for instrument operation or to maintain an inert atmosphere, and energy: steam or Dowtherm for heating, refrigeration, or electricity for power.

Plant economics depends on procedures that must be carefully engineered to render the process attractive. Equipment size is selected to optimize capital investment. The operating cycle is chosen to reduce equipment idle time. Cycle scheduling is programmed to minimize operating costs. The plant may run for the whole year, or activities may be reduced to an optimum operating period beyond which the equipment is idle. Depending on whether one or several products are made in the same plant, a batch processing plant may be one of the following types.

TABLE I Batch Polymerization Cycles^a

Batch cycle	Cycle time (hr)	
	Cycle 1	Cycle 2
Materials charge	0.5	0.5
Preheating	1.0	0.0
Polymerization	8.0	4.0
Stripping unreacted	1.0	0.5
Cooling	1.0	0.0
Discharge	0.5	0.5
Reactor flushing and cleaning	1.0	0.2
Maintenance	0.5	0.5
Inefficiency (time losses)	0.5	0.3
Total cycle time	14.0	6.5

^a Reactor volume, 3700 gal; load, 5300 lb.

A. Single-Product Plant

When production volume is sufficient, it is economical to build one plant for one product. Batch production in a single unit may be limited by maximum reactor size. Holdups of greater than 20,000 gal are handled in separate parallel reactors. To use common upstream and downstream facilities, the reactors may not be operated simultaneously but on overlapping schedules. When long reaction times cannot be avoided, the reaction sections operate batchwise; however, feeding reactants and recovering products may be continuous for economic reasons. This practice is typical of many processes, such as the saponification of natural fats in intermediate quantities. In the production of ethanol by fermentation, two reactions (saccharification and fermentation) are operated on a batch basis, while hydrolysis (conversion of starch to dextrin) and product recovery by distillation are continuous.

The interaction between the batch and the continuous parts of the plant is a major concern of plant design and operating procedures. When the production volume reaches certain levels, continuous operation may be more economical. More often, batch chemical plants are built to produce more than one product. The objective is the use of existing equipment, which becomes idle when market demands are fulfilled. These plants can be classed in one of two broad categories: multiproduct and multipurpose. The difference between the two types is illustrated in Fig. 1.

B. Multiproduct Plant

A given battery of equipment is used to produce only one product at a time. The same equipment is used for the manufacture of different products by changing operating pro-

TABLE II Block Cycle of a Multiproduct Plant

Cycles	Product	Total time (hr)
24	A-1	61
19	A-2	49
110	A-3	280
27	A-4	69
4	<u>Downtime</u>	<u>32</u>
180	All 4	491

cedure. The products handled are chemically similar, and product changeovers are relatively simple. Product quality can be altered; new products or product grades can be tried, and product capacity can be adjusted to meet market demand by scheduling changes rather than by changing equipment. Several products can be scheduled for production in sets of various cycles, one set following the other. The batch cycle of the plant takes into account the various products handled by the same plant. Each product has its own batch cycle. The block cycle is the time needed for the once-through manufacture of all the products at the annual production rate. For illustration, consider a multiproduct plant that makes 100 million pounds per year of four products: 13.5% A-1, 10.5% A-2, 61.5% A-3, and 14.5% A-4. The plant makes only one product at a time in batch cycles of ~153 min. The front part handles the reaction on a batch basis. A second reaction and product recovery take place continuously in the back end of the plant. Since three products must be stored while the fourth one is manufactured, bulk storage must be provided for four products and their raw materials. A typical block cycle for the four products consists of 491 hr distributed in 180 cycles (Table II). Each product cycle includes reactants loading, batch reaction, residual reactant stripping, product transfer, and downtime between batches. Eight hours of downtime and cleanup are allowed to switch from one product to the next. The number of product cycles per block cycle is set from market demands. The size and the capital cost of the bulk storage needed increase in proportion to the block cycle time. The effective operating time decreases when the block cycle time is reduced, due to the downtime needed to switch from one product to the next. Hence, the size and the capital cost of the process equipment (battery limits) must increase to make the same annual production rate. There is an optimum block cycle that must be selected to minimize capital costs of storage and battery limits. The optimum is between 3 and 5 weeks; it is not largely dependent on downtime and cleanup between products nor on the extra capacity of product storage, which may be needed to ensure enough raw material supply.

C. Multipurpose Plant

A battery of process equipment is used in various arrangements to manufacture more than one product at one time.

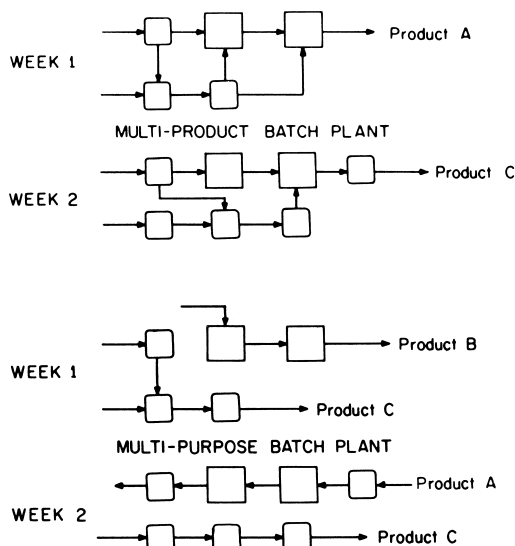


FIGURE 1 Multiproduct batch plant (top) and multipurpose batch plant (bottom).

The goal is to use common equipment in the manufacture of various products needed by the market via production planning and scheduling. An individual route may be selected for each batch of product, depending on the availability of the various pieces of equipment. The products may have low production requirements or may be at the beginning of their commercial life. These plants are rare, since when they are constituted, they are likely to be re-defined as single-product plants for economic reasons.

D. Batch Process Development

Conception is the most critical step in the development of a new process. Though still done largely on the basis of experience and intuition, it may be implemented with process synthesis. Computerized algorithms may provide for a large number of possible routes to a product. This method, combined with the analysis of raw material costs and DS-51 ASTM tests for process hazardousness, are the best options to speed up the bench-scale development of new chemical processes. Thus, one or a reduced number of routes to the desired product can be identified for preliminary process development.

Specification of a chemical route and processing steps rarely defines the process completely. The process designer is still free to choose appropriate design variables. Further difficulties are encountered due to the lack of commercial or engineering data needed for the design.

Preliminary design, assisted by material and energy balances, contributes to flow-sheet development and to establishing process operating conditions. Also, it serves the purpose of equipment sizing and generating a list of equipment for preliminary process evaluation. The first capital investment estimate of the process is based on this study. Then, a pilot plant test of the new process can be scheduled. Product samples are made for customer analysis to set product specifications. Physicochemical and engineering parameters are measured in the pilot plant to verify preliminary educated guesses of physical properties and transport parameters which need to be firmed for the definitive design.

Economic evaluation analyzes the process on the basis of production costs and the rate of return on capital investment. If the process is economically attractive, the construction of a commercial plant is recommended to top management. A definitive design package is generated for construction engineering. A definitive mechanical flow diagram is included, with all the electrical, instrumentation, and safety needs of the process. The waste disposal and utilities sections are completed. The design engineers must ensure that the process will run safely and it will be environmentally clean. Several process and plant permits, without which plant construction cannot proceed, need to be requested at this time.

When construction is completed, design, plant operations, and the construction group proceed to start up the new plant. Technical service takes over to operate the plant to the best economic advantage. Any design errors are overcome with production planning strategies and short-term production scheduling of the various products that will be produced in the new plant. The time invested in the development and construction of the new process is nearly four years. Throughout all this time, the opportunity for formal optimization is minor because the new process is not fully defined. In the course of chemical business, the next step for the plant is debottlenecking and production expansion planning.

V. ECONOMICS OF BATCH OPERATIONS

Batch plant economics are similar to those ruling continuous plants: They are set by the revenue collected and the total cost of running the plant at a certain production rate. The analysis of the cost of manufacturing one or more products is often based on the F.O.B. cost of the process and storage equipment required; the control equipment needed to run the plant; the cost of additional materials and labor needed to construct the plant; the indirect costs, including supervision as well as engineering and construction fees; and additional costs due to freight, insurance, and taxes. A list of the process and storage equipment results from a preliminary design of the plant, which also yields a process flow diagram (Fig. 2). The equipment list is used to make a capital investment estimate of the plant, which includes the total F.O.B. cost and all the other expenses of constructing a plant. When real cost data are missing, one can approximate values using Guthrie's standard procedures and extensions for batch plants. This analysis yields the total capital costs for the process at the production rate at which the plant can operate. In definitive cost analysis a list of control equipment is produced with the completion of a definitive design package, which leads to a mechanical flow diagram. The total book manufacturing cost of the product is a composite of four costs: (1) raw materials and utilities, which are proportional to production rate; (2) direct production costs, which include (a) operation labor and overhead (supervision and clerical work), (b) operating supplies, which are proportional to operating labor, (c) maintenance and repair, which include labor and materials (these costs are proportional to fixed capital costs), and (d) laboratory labor; (3) plant overhead, proportional to operating plus maintenance and repair labor; (4) fixed charges, which include depreciation proportional to total capital costs, as well as local taxes and insurance proportional to fixed capital costs.

Total product cost is obtained by adding to the book manufacturing costs the following costs: (1) general

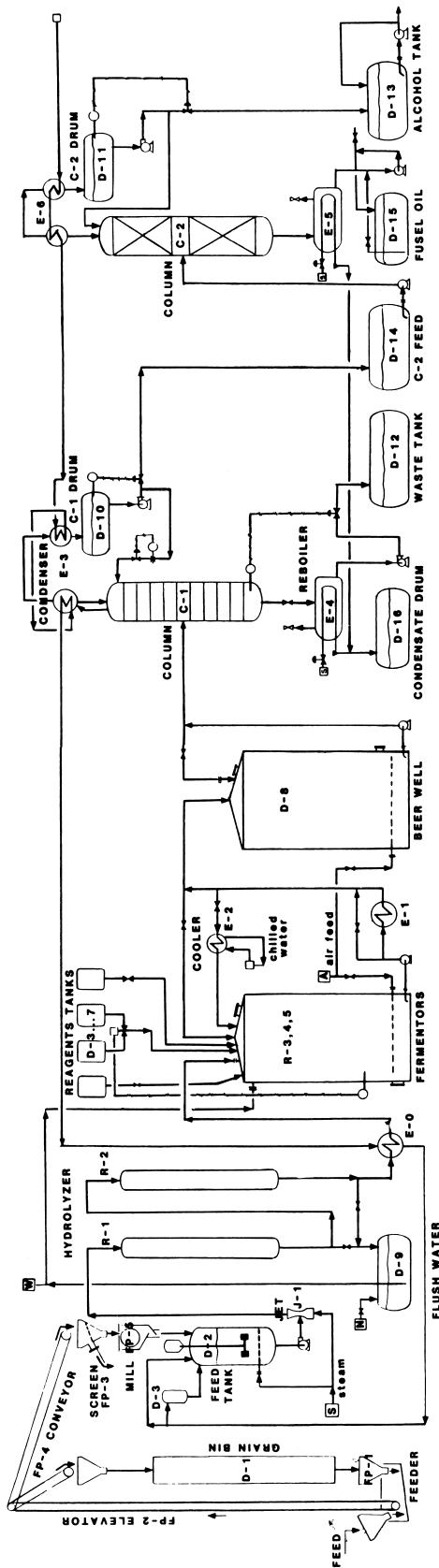


FIGURE 2 Process flow diagram: manufacture of alcohol fuel.

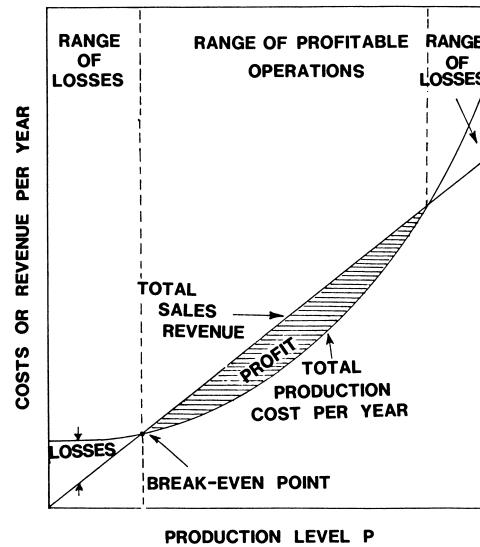


FIGURE 3 Break-even analysis chart.

administration costs and expenses to sell the product, (2) freight cost of delivering the product to the distributor, and (3) markup, which must include the recovery of the capital invested and the interest.

Sales revenue is the total amount collected from selling the plant production P (pounds/year) at the selling price of A dollars per pound of product. This presumes that all the product made can be sold. Product selling price is determined by market conditions.

A. Break-Even Analysis

Selling price may change depending on the production level P due to market circumstances. Figure 3 shows a plot of the sales revenue, which, for illustration, is linear with P . It also shows the total product cost plotted at various production levels, which nearly depends on the 0.6 power of P . At high production levels, total production costs lay below the possible sales revenue; the difference between the two costs indicates the total profit at that production level. At low production levels total production costs are larger than the possible sales revenue, and plant operation yields losses. The production level at which total costs equal the possible revenues is the break-even point. Plant production and sales must run at higher rates for the operation to be competitive.

B. Batch Cost Minimization

The production capacity of a process is determined by market demands. When the break-even analysis of a process flow scheme indicates losses at the maximum marketing level, its production capacity cannot be increased to reduce product cost. The process design engineer must reduce

capital investment to a level which will render economical a process for a single-product plant. One or several trains of equipment may have to be replaced by less expensive trains. The new process is expected to be safe and environmentally clean and to meet proper product specifications. The utilization of standard, used, or idle equipment may bring an economic solution to the design.

When a single-product scheme cannot be sustained, the cost of capital investment may be split among several products. A multiproduct or a multipurpose plant may be designed at a minimum capital cost for the yearly production rates needed. This procedure may lead to a lower product cost. Raw material and operating costs can be minimized simultaneously. Also the operation of the multiplant may be scheduled to attain the minimum operating costs of the various products handled.

A cost equation may be written to include all the costs, which are expressed in terms of the capacity of the flow-scheme components. The selection of equipment sizes that minimize capital investment is (1) complicated by interrelations between pieces of equipment, (2) limited by the discontinuity in size of standard equipment, (3) fixed by the availability of idle or used equipment, and (4) restricted by the higher cost of custom-made equipment. Writing one equation for a complete plant is a complex task. It is more likely that it may be done for small sections of the plant which can be operated as interrelated trains.

The cost equations thus written are discontinuous functions of the size of the units which compose the trains. A mathematical minimization of any of these equations may not lead to a practical minimum. It may indicate only the domain where less-expensive solutions may exist. The practical alternative is to draw flow schemes which are equivalent to the process under investigation. The economic analysis of these schemes terminates with the selection of one which requires the minimum capital investment and operating costs.

C. Batch Process Management

In a multiproduct plant, all products follow nearly the same path through the process. Only one product, or in a complex case a few products, is produced at one time. Although it is feasible to alternate batches of different products to reduce idle time of equipment, this is rarely, if ever, done to minimize operational error and cross-contamination. Product changeovers are not frequent, especially if extensive cleanout between changeovers is required.

The operating decisions to be made for a multiproduct plant are more complex than for a single-product plant. The useful production time of making several products is diminished by the time for overlapping product changeovers. Operation without product overlap results

in less useful time, since a new product can be started only when the complete plant is empty and clean. It is convenient to produce only one product in a given week. The number of batches produced for each product depend on the time allocated to the product. In this task, computerization may help to minimize storage of raw materials and reduce the cost of working capital, while customer orders are filled on time.

Multipurpose plants are adequate for the simultaneous manufacture of many products on a small scale. This is typical of the pharmaceutical or specialty chemicals industries. A structure housing general-purpose equipment can be set up and used as required. The trains available should be sufficient so that a large number of products can be produced at a time. Their size—small rather than large—should be adequate for the production volumes of the products handled. If it is economical, one product may have more than one production route and procedure. The various products compete for raw materials, utilities, manpower, and production facilities. With no interaction between products, the plant may operate with overlapping or nonoverlapping cycles and product changeovers. When products interact, the situation becomes extremely complex.

Production planning studies the amounts of the various products which should be manufactured in a given plant when the production requirements are known. This problem must be considered during the plant design stage. Scheduling decides in which equipment each product should be manufactured, when the operation starts, and when it is expected to finish.

Production planning and scheduling problems arise in multiproduct plants because the production time available on each train must be allocated to the manufacture of a large number of products within certain production constraints. Short-term industrial scheduling is a complex problem because the day-to-day plant status, customer needs, raw materials, and personnel availability are changing continuously. The demand for computer-aided scheduling tools is increasing because production schedulers are linked to their sales and service personnel through computerized reports on customer orders, inventory levels, and plant status. Using such information would reduce the complexity of the scheduling job, speed up operations, and produce more economical selections.

Production planning problems result from substantial changes in market demands, the need to introduce new processes, and the increasing pressure on existing production capacity. On these accounts the operating plan may require radical changes in the use of equipment. A computerized algorithm, BATCHMAN is useful in determining the best equipment configuration for the optimal product combinations and assigns the available equipment

to tasks on the basis of achieving the best performance. That program could be the starting point for a more detailed short-term planning analysis.

VI. OPTIMIZATION

Production costs depend on the cost of the manufacturing equipment involved. The batch times $t(i)$ determine the sizes of the batch units; the longer these times, the larger is the holdup storage needed to keep the stock of reactants and products used in the next batch. The semicontinuous times $t(j)$ may be reduced to decrease the total batch cycle, at the expense of increasing the size and the cost of the semicontinuous units. The reduction of storage and reactor sizes may result in an economic advantage. Studies of this sort are completed during the design of the plant to minimize capital investment. Any reduction in the downtimes $t(k)$ also improves the economics. Though this analysis may seem simple, it indicates the importance of considering the specific circumstances ruling batch processes. Every case is different and may involve additional complexity. For example, production costs could be reduced by decreasing cycle time if the operation did not result in additional costs caused by secondary unfavorable reactions. It can be inferred from the above considerations that in batch processing there are opportunities for the analysis of alternatives that may well pay back the time and efforts invested. Some concepts of optimal design may be useful, recognizing that formal optimization may be too expensive; that is, a generalized expression for batch plant optimization problems may not be applicable to actual cases, since the cases to be analyzed are too specific.

VII. SIMULATION

Experience has been the only requirement for running many processes. However, there is economic incentive for modeling and simulating industrial processes since, when technology is understood on some theoretical basis, it is likely that process operating efficiency and economic yields will be improved. Modeling of batch processes has attracted considerable attention because of the need to determine the changes occurring and to follow them with time. A model consists of a set of mathematical equations and proper boundary conditions that are capable of simulating a plant or a section of a plant. It can also be used for process prediction and control. A model must be simple enough to be understood; it must be suitable for predicting the behavior of the system it intends to simulate; and it must not be trivial to the extent that its predictions are grossly inaccurate. The first step in the theoretical analysis is understanding the physicochemical principles that rule the transformations taking place. Then the mathe-

matical description of the process can be formulated. At this stage, the equations do not constitute a model; they only interpret the principles in mathematical form. Next, the important characteristics of the problem are identified, and the mathematical formulation is simplified, justifying the approximations with physical arguments. The resulting equations constitute a model. Though the description of large-scale processes is complex, the model postulated must contain the essence of the process in its simplest form; it must be expressed by a set of equations whose solution can be obtained and be useful for process simulation. Complications can be added to simple models as needed; beginning with a model that is too complex causes confusion and unnecessary work. The depth of physical knowledge determines the type of modeling that can be formulated; two types are common, as described in the following subsections.

A. Defined Modeling

Defined modeling is applicable to processes where one can associate sufficient measurable inputs and outputs with each physical change taking place in the system. In most cases, the inputs and outputs can be related through a set of nonlinear ordinary (ODEs) or partial differential (PDEs) equations. Modeling is the determination of the equations and boundary conditions that define the system in space and time. Simulation results from the solution and numerical evaluation of the model. The equations are set by establishing mass and energy balance and/or applying other pertinent physical laws. Modeling problems of this type are common in engineering practice. A typical case is that of modeling a batch chemical reactor. Then one material balance equation per component is required in addition to defining the inventory change of each component and the energy with time. Two varieties of defined modeling are common:

1. *Transport phenomena modeling.* This type of modeling is applicable when the process is well understood and quantification is possible using physical laws such as the heat, momentum, or diffusion transport equations or others. These cases can be analyzed with principles of transport phenomena and the laws governing the physicochemical changes of matter. Transport phenomena models apply to many cases of heat conduction or mass diffusion or to the flow of fluids under laminar flow conditions. Equivalent principles can be used for other problems, such as the mathematical theory of elasticity for the analysis of mechanical, thermal, or pressure stress and strain in beams, plates, or solids.
2. *Empirical modeling.* When the model is too complex or some defined knowledge is missing, approximations of theory are used to implement the theoretical arguments.

This is done in simulating complex reactions such as polymerization. The zeroth, first, and second moments of the molecular weight distribution are used to characterize the kinetics of the reaction.

B. Stochastic Modeling

Stochastic modeling is used when a measurable output is available but the inputs or causes are unknown or cannot be described in a simple fashion. The “black-box” approach is used. The model is determined from past input and output data. An example is the description of incomplete mixing in a stirred tank reactor, which is done in terms of contributions of dead zones and short circuiting. In these cases, a sequence of output called a *time series* is known, but the inputs or causes are numerous and not known; in addition, they may be unobservable. Though the causes for the response of the system are unknown, the development of a model is important to gain understanding of the process, which may be used for future planning.

Two modeling methods are current. The first is determining the parameters of a model of the form:

$$y(k) = f(k) + w(k), \quad k = 1, 2, \dots, N \quad (2)$$

where $f(k)$ is a given function and $w(k)$ represents a random noise sequence. The objective is to select the parameters in such a manner that $w(k)$ is a zero-mean white noise sequence of minimum possible variance. The common forms of $f(k)$ are a finite power series in k or a simple exponential. The second modeling method involves considering that the time series is a linear transformation of a zero-mean white noise sequence. In simple cases, this may be regarded as the output of a linear, time-invariant, discrete-time system subject to white noise input, and the object of the modeling is to estimate the parameters of the transfer function or the difference equation of the system. These are recognized as stationary time series models. When a model of this type is not adequate, it may be necessary to fit a difference equation model to the n th successive differences of the given time series; this is called a *nonstationary time series model*.

C. System Modeling

Mathematical modeling of systems for which characteristic variables are time-dependent only and not space-dependent is done by ordinary differential equations (ODEs). The situation is found in a nearly well-mixed batch reactor. There one may find differences in temperature or concentrations from one site to another due to imperfect mixing. When space changes are not important to the model, the process variables can be approximated by means of lumped parameter models (LPMs). When the

space change of certain variables is important, the description of these process variables must be in terms of partial differential equations (PDEs). The changes of these variables may be interpreted only by distributed parameter models (DPMs). An illustration is encountered in a system of large viscosity where the temperature changes along the direction of heat flow are as important as the time changes in determining product quality.

Algebraic equations (AEs) describe relations among variables of the system which are independent of time or space changes. They represent variables not related by material energy or momentum balances. They may characterize physicochemical or other type of relationships between physically independent portions of the system, as the vapor and liquid spaces of a reaction vessel.

D. Single-Stage Modeling

The model for a single unit or a train of equipment which is a part of a plant may consist of one or two PDEs, combined with a few ODEs and some AEs. The mathematical solution of a model having more than two PDEs may be beyond the budget of many industrial organizations. The PDEs stand for a temperature and a concentration or another variable whose variations in space are as important as their changes in time for the proper characterization of the system (i.e., for determining product quality). The ODEs represent the kinetic or dynamic changes of the unit with time, and the AEs indicate the equilibrium relationships between the various phases present.

Although the model is only for a single unit, it is by no means simple. Its solution may be pictured as a plot which shows the changes of the mean concentrations of the reactants, some intermediates, and the products with time during a batch operation. It will also indicate the space distribution at various times of the temperature and the particular concentration which characterizes the product quality.

A single-stage model is the simplest model which can be proposed to simulate a single unit or one train of equipment that is part of a plant. The model represents the operation of units or pieces of equipment which must run for the same length of time.

E. Multistage Modeling

Since a model of just a single unit of a plant is by itself complex, modeling of a complete plant, no matter how simple it is, cannot be achieved unless the plant is split into sections which are relatively independent from each other and small enough to be represented by single-stage models. Thus, plant modeling is converted into a collection of single-stage models of which several may be handled

simultaneously, if they happen at the same time, or in series as the time for them to happen arrives.

VIII. ANALYSIS

The importance of modeling batch processing systems forces a review of the mathematical analysis needed to set up and solve the models. The mathematical definition of physical problems involves: (1) identification, (2) expression of the problem in mathematical language, (3) finding a solution, and (4) evaluating the solution. The completion of these steps in the order established determines whether a solution can be attained. The problem must be identified before one spends time setting up equations; these and the initial and boundary conditions that define the problem must be well established before a solution is attempted; then a solution can be obtained and evaluated.

A. Problem Identification

Problem identification is a human engineering problem; it arises from discussions with individuals working where the problem exists. The problem solver must listen to those individuals, use data to identify the problem, set up a model that best approximates it, fit the model to the data, solve it, and compare the solution with the actual situation. Then, the problem solver must use the results and capitalize on any learning from the completed analysis. Problem identification leads to system definition, which is best accomplished by using a basis of thermodynamics. Physical boundaries are set that separate the system from its surroundings. A system may be isolated, closed, or open. Isolated systems do not exchange energy or matter with their surroundings. The energy E and the mass m of the system remain constant during the process: $dE = 0$; $dm = 0$. We deal with systems that change with time; hence, any changes in energy or mass must occur within the system. Consider a system composed of vapor and liquid; a third phase (solid), added at a given moment, reacts with the liquid. The three-phase system is isolated, but there is energy and mass exchange between the solid, liquid, and vapor. As the solid dissolves and reacts with the liquid, it produces heat, which vaporizes some of the liquid, but no energy or mass is lost outside the three phases. Closed systems exchange energy, but not matter, with their surroundings. The total mass of the system is constant, as it is for an isolated system. The change in energy is given by the law of energy conservation,

$$dE = \delta Q + \delta W \quad (3)$$

The term dE in Eq. (3) is an exact differential; that is, its value is independent of the path followed by the system during the change; it depends only on the initial and final

states of the system; the terms δQ and δW are inexact differentials since their values depend on the path followed by the system during the change. Open systems exchange both matter and energy with their surroundings. If $\delta\psi$ is the resultant energy intake during the time dt due to heat transfer and exchange of matter, the changes of energy and matter of the system are

$$dE = \delta\psi + \delta W \quad \text{and} \quad dm/dt \neq 0 \quad (4)$$

1. Constraints

Problem definition requires specification of the initial state of the system and boundary conditions, which are mathematical constraints describing the physical situation at the boundaries. These may be thermal energy, momentum, or other types of restrictions at the geometric boundaries. The system is determined when one boundary condition is known for each first partial derivative, two boundary conditions for each second partial derivative, and so on. In a plate heated from ambient temperature to 1200°F, the temperature distribution in the plate is determined by the heat equation $\partial T/\partial t = \alpha \nabla^2 T$. The initial condition is $T = 60^\circ\text{F}$ at $t = 0$, all over the plate. The boundary conditions indicate how heat is applied to the plate at the various edges: $y = 0$, $0 < x < a$, $\partial T/\partial y = 0$; $y = b$, $0 < x < a$, $\partial T/\partial y = 0$; $x = 0$, $0 < y < b$, $\partial T/\partial x = 0$; $x = a$, $0 < y < b$, $-k(\partial T/\partial x) = h(T - T_A)$. The first three conditions indicate that the plate is insulated while it is heated by convection along the fourth edge, $x = a$, from an environment at temperature T_A .

B. Mathematical Expression

Batch processing problems are described in terms of one or more differential equations, sometimes combined with algebraic or integro-differential equations. The equations indicate changes in the system with time along its geometry. The geometry is defined by means of a standard set of space coordinates: x, y, z , Cartesian; r, θ, z , cylindrical; r, θ, φ , spherical; or other type. System properties such as temperature, concentration, and velocity may change in both time and space. Most real situations are described by nonlinear equations, which cannot be solved by analytical means. The popularization of computers and the development of numerical integrators, such as GEAR, EPISODE, DASSL, and others which can handle almost any situation, has made possible the solution of many problems defined in terms of ODEs both combined and not combined with algebraic equations. A good number of cases that are set in terms of PDEs can be handled by transformation of the PDE to sets of simultaneous ODEs. The limits of the solution are determined by the number of simultaneous equations generated, the size of the computer,

and the computing time that can be spent solving each case. The formulation of the equations that represent the physical situation is not highly complex, but it cannot be reduced to a routine procedure. The equations result from the application of basic laws of physics to the various cases. Commonly used for this purpose are the laws of conservation of mass, energy, and momentum and several others. The problem is determined by defining the system being considered, the boundary conditions affecting the system, and a balance of mass, energy, or momentum (MEM), rates expressed by:

$$I - O + G = A \quad (5)$$

where I is input, O output, G generation, and A accumulation of MEM in the system. In certain situations the MEM balance may have to be implemented or replaced by equivalent physical laws. Thus, in a problem of fluid mechanics, the basic elements of analysis are stresses and strain rates. A force balance generates the equations of motion; a mass balance yields the equation of continuity. These balances must be implemented with the relationship between stresses and strain rates, which must be established experimentally. When the classical approach of balancing MEM rates [Eq. (5)] is not possible due to complexity, the system may have to be redefined in simpler terms, or a time history of the process may have to be used in its definition.

C. Solution of the Problem

The essence of solving the problem is shown in Fig. 4. There are two ways in which the basic equations can be solved: by numerical means and by analytical procedures. In general, the PDEs or ODEs that describe actual situations are nonlinear and must be solved numerically using a computer. Each PDE is transformed into a set of ODEs by the method of lines. The ODEs are reduced to the solution of initial value problems,

$$\begin{aligned} dY/dt &= f(Y, t), \\ Y &= Y_0 \quad \text{at} \quad t = 0 \end{aligned} \quad (6)$$

where $f(Y, t)$ can be any piecewise continuous linear or nonlinear function that can be integrated numerically in time t . The problem is solved by numerical integration. Two types of integrators are available for the purpose: (1) nonstiff integrators, the traditional ones of which use the Runge–Kutta method and the more efficient of which use the Adams–Bashforth–Moulton predictor-corrector method; and (2) stiff integrators, which are based on GEAR stiff formulas, implemented by Hindmarsh–Byrne. They are effective for fast integration of very stiff problems. Two sets of stiff integrators are available to users at

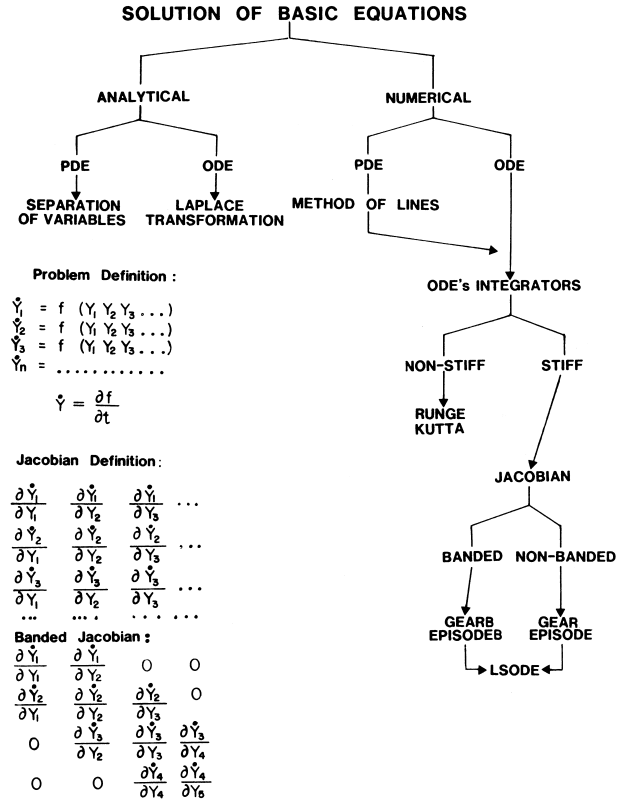


FIGURE 4 Methods of problem solution.

a nominal cost: (1) nonbanded Jacobian integrators such as GEAR and EPISODE, which are designed to handle sets of equations of the form of Eq. (6), whose Jacobian consists mostly of nonzero elements all over the matrix; and (2) banded Jacobian integrators such as GEARB and EPISODEB, which are suitable for handling sets of equations of the form of Eq. (6) possessing banded Jacobians, as in the case of tridiagonal or pentadiagonal matrices or similar matrices that contain zeros in one or more lines parallel to the diagonal.

Although the solution of the problem must be accomplished by numerical methods, analytical solutions are of extreme value. Numerical integrators must be tested using certain known solutions to ensure that the algorithm is free of programming errors and that the approximations invoked in the numerical procedure do not generate unacceptable arithmetic errors. Analytical procedures are applicable to the solution of linear problems defined in terms of PDEs. Among these are the method of separation of variables and the method of Laplace transformations. The use of these methods may require intervention of a skilled user who can resolve the complexities arising in the applications. The difficulties encountered in applying the first method are in the evaluation of the integration constants, which require the identification of functions

orthogonal to the functions being integrated. The problems found in the use of the second method are in the inversion of the final transform, which leads to the solution of the PDE.

D. Evaluation of the Solutions

Numerical evaluation of the solutions to these problems is not easy. Most analytical solutions to PDEs are expressed in terms of infinite series. To get numerical answers one must evaluate series functions, which often are not sharply convergent; thus, the number of terms to be accounted for in the summation may be very large, or the accuracy of the answer may be largely reduced. Alternatively, when the solution is obtained by numerical methods, the accuracy is affected by the approximations made in replacing partial derivatives by finite difference expressions. Then special consideration must be given to estimating the truncation, roundoff, and generated errors introduced by such approximations. Another problem inherent in numerical solutions is that there is no way of telling whether the answer obtained is correct. It is customary to solve the problem independently by a different numerical method and compare the answers to check for errors due to the other numerical method. However, the accuracy of the answers cannot be estimated except by comparison with analytical solutions of particular cases. Therefore, in evaluating the solution, it is important to give consideration to the following topics: (1) analysis of errors resulting from numerical approximations, (2) comparison of numerical solutions to check the correctness of the software, and (3) solution of analytical approximations to verify the validity of numerical solutions.

E. How Good is the Model?

A mathematical model of a plant or a section of a plant can be judged only by comparison with actual plant data. The model may be considered as good when the simulated variables can predict with some level of confidence the plant parameters which are important in determining the cost and quality of the finished product. Failures of the model are likely to be a result of: (1) oversimplification of the equations that constitute the model, (2) inadequacy of the numerical solution of the equations.

The solution to the first problem is limited by the increase in time or the computer capacity available to solve more complete or more advanced equations. The second problem is even more difficult to acknowledge. It may be due to error accumulation through the nonlinear domain. The numerical solution of a differential equation is based on the approximation of time and, in the case of PDEs, space partial derivatives, by finite-difference equivalents.

Because of the accumulation of these errors, in time or space, the numerical results which are generated by simulation may be far from the true solution. This problem may be recognized by comparing two solutions of the same problem using 100 times smaller time or space increments, or both, when it is the case. Solutions passing that test are likely to be correct. But such a test may be too stringent in a good number of cases since it may require too much computing time or too much computer memory to be economical to run.

It is proper to emphasize the fact that there is no mathematical way of determining how correct is the numerical solution of a nonlinear differential equation, just by looking at what an algorithm predicts for linear forms of the same problem. Even if a computer program generates accurate numerical solutions which are in agreement with analytical solutions of the linear forms of a nonlinear equation, the solutions in the nonlinear domain may not be correct.

IX. SAFETY

In batch plants, as in continuous plants, hazardous materials are stored or handled. The operating conditions of continuous plants are set to completely avoid fire and explosions. Risks are taken only in the startup or shutdown of the plant, when process leaks occur or undesired materials are released for protection of the equipment. In batch plants the safety risks are more numerous, since equipment and materials are brought more often in contact with ambient conditions or are subjected periodically to transient temperature, pressure, or concentration conditions, which may constitute safety hazards. Therefore, safety is much more important in the successful management and operation of batch plants than in that of continuous processing plants. Safety must be a major consideration in the storage and handling of volatile or chemically unstable reagents or intermediates. Equipment selection and design must consider the vapor pressure and the thermal stability of the reagents and protection of the equipment against excessive pressure due to process conditions, to runaway reactions, and to emergency situations arising in a fire or explosion. Process equipment is designed to withstand internal pressure or vacuum and must meet the requirements of Section VIII of the ASME (American Society of Mechanical Engineers) code. Storage equipment usually withstands very low internal pressure or vacuum and is constructed to meet the regulations of the ASME and the API (American Petroleum Institute) codes for storage vessels. Safety and emergency relief equipment and design must fulfill the requirements of the NFPA (National Fire Protection Association) codes and procedures. Safety

and health protection of the personnel in charge of the plant must meet OSHA (Occupational Safety and Health Administration) regulations. The protection of the environment against chemical contamination must meet national, state, and local EPA (Environmental Protection Agency) regulations. Protection of the capital investment in the plant and the equipment in the case of fire or hazardous conditions must meet the codes of the National Board of Insurance Underwriters. Technical recommendations for the proper design of an emergency relief system (ERS) have been compiled by the Design Institute for Emergency Relief Systems (DIERS) of the American Institute of Chemical Engineers.

Technology leading to the safe design of industrial processes is available. ASTM offers a computerized procedure, DS-51 (1974), to test the reactants, intermediates, and products of a process early in its development. The program CHETAH rates the components of a process as hazardous if they are likely to undergo decompositions which may result in thermal reaction runaways. The rating is based on the energy release due to probable oxidation of the atoms in the molecule at higher temperatures. These situations may result from a process upset, a fire situation, or by an explosion occurring in the reactor.

Molecules containing O, S, P, N, F, Cl, Br, and I are chemically unstable and should be suspicious. If any of the components of a process are rated hazardous, the process may be substituted for one which uses no such components, or experimental tests for the safe design of the pilot plant and the commercial process must be scheduled.

The safe design of a process which uses hazardous chemicals requires testing the reacting mixture to measure the rate of chemical decomposition during any reaction runaway. When the components decompose in several stages, the experimental study must be carried out to high temperatures to ensure the complete decomposition of all the components of the system. Multistage decomposition is typical of chemicals containing two or more of the atoms listed above, such as halogenated compounds which oxidize one halogen atom at one temperature, the second at a higher temperature, and so forth. Completeness of the decomposition must be verified via a material balance and chemical analysis of the products made and the residue remaining in the decomposition autoclave.

Explosions resulting from process upsets are detonations or deflagrations. Processes complicated with detonations cannot be used commercially since the rates of pressure rise during the upset are too large (about 100,000 psia/sec) to be relieved by any existing technology. Industrial deflagrations are thermal runaways where the rates of pressure rise are about 4000 psia/sec or less. They are due to reactions with combustion-supporting materials such as oxygen, or any of the other elements cited

above. These runaways may be relieved safely. Chemical reactions which degenerate in deflagrations characterized by pressure-rise rates higher than 4000 psia/sec may not be relieved safely for lack of suitable industrial technology and should not be used for industrial purposes.

A bursting disk is used to protect the reaction vessel. It should be properly rated to ensure that the internal pressure developed during the upset does not exceed the allowance made by ASTM Code Section VIII (or any other pertinent section) for pressure vessels. The actual disk relief area used is 25 to 60% larger than indicated by process requirements (as reported by the disk manufacturer) to account for incomplete bursting of the disk. An emergency relief system (ERS) manifold should be used to collect and convey the effluent material to a train of equipment where it is recovered or properly destroyed to avoid environmental contamination.

The DIERS Institute of the American Institute of Chemical Engineers has developed procedures for the safe ERS design of processes undergoing thermal runaways caused by deflagrations. Runaways may be of three types:

1. *Vapor systems*, where boiling is reached before potential gaseous decomposition. The heat of reaction is removed by vaporization of the solvent present or added on purpose to keep the system thermally stable.
2. *Gassy systems*, where a gaseous decomposition occurs in the absence of tempering. The total pressure developed during the upset is due to the presence of noncondensable gases.
3. *Hybrid systems*, where gaseous decomposition occurs before reaching boiling, but the rate of reaction is tempered by vapor stripping. The pressure developed in the system is due to the vapor pressure of the volatile components and to the partial pressure of noncondensable gases.

There are two approaches to ERS design. One is system modeling, which identifies the cause of a pressure rise from a hazard analysis. It uses approximate models—all-vapor flow, all-liquid flow, or two-phase flow—to simulate the pressure increase of the reacting system vs. time and to determine vent size. The method is complex since it must identify the stoichiometry, the mechanism, and the kinetics of the decomposition causing the pressure rise. Two pressure models are used for vent sizing:

1. *Low-pressure models* applicable to the protection of process buildings and storage tanks ruled by the API codes. RUST's low-pressure model is usually successful for vent design of internal or external overpressure.
2. *High-pressure models* applicable to pressure vessels and chemical reactors subjected to more than 15 psig

of internal pressure. Two computer programs, SAFIRE and DEERS, were developed by the Design Institute for Emergency Relief Systems (now a branch of the Center for Chemical Process Safety of the American Institute of Chemical Engineers) to rate the relief area needed for safe design of high-pressure vessels. The validity of the results from these two simulation programs depends on the assumptions made for the critical flow of two-phase flow systems through the exhaust manifold. Unfortunately, the hydrodynamics, and especially the pressure drop of gas-liquid-solid systems flowing through a manifold are not well known. Also, the assignment of liquid and solid phase entrainments to the vapor phase outflowing the vented vessel may not be realistic.

The simpler and most reliable approach to the use of the DIERS methodology is the use of FAUSKY's reactive system screening tool (RSST). It is an experimental autoclave which simulates actual situations that may arise in industrial systems. The RSST runs as a differential scanning calorimeter that may operate as a vent-sizing unit where data can readily be obtained and can be applied to full-scale process conditions. The unit is computerized and records plots of pressure vs. temperature, temperature vs. time, pressure vs. time, and the rates of temperature rise and pressure rise vs. the inverse of temperature. From these data it determines the potential for runaway reactions and measures the rates of temperature and pressure increases to allow reliable determinations of the energy and gas release rates. This information can be combined with simplified analytical tools to assess reactor vent size requirements. The cost of setting up a unit of this kind is close to \$15,000.

X. MATERIALS

A. Properties of Materials

Elasticity of solids determines their strain response to stress. Small elastic changes produce proportional, recoverable strains. The coefficient of proportionality is the modulus of elasticity, which varies with the mode of deformation. In axial tension, E is Young's modulus; for changes in shape, G is the shear modulus; for changes in volume, B is the bulk modulus. For isotropic solids, the three moduli are interrelated by Poisson's ratio, the ratio of transverse to longitudinal strain under axial load.

When solids deform almost to the breaking point, they exhibit brittle behavior; the stress at fracture is several orders of magnitude lower than the computed strength. The loss is ascribed to the presence of minute cracks probably formed during solidification. Compressive stress can induce crack propagation; the magnitude of the stress at

fracture is almost ten times larger in compression than in tension. In dealing with large strains, one must distinguish between conventional stress, which is the axial load divided by the original cross-sectional area, and true stress, which is the load divided by the actual cross-sectional area.

A slip or glide of part of one body over the other results in plastic deformation. At the beginning of plastic deformation, the stress produces a permanent strain on the material. The progress of plastic deformation in those materials exhibiting strain hardening is marked by strain hardening; each additional increase in deformation requires an additional increment of stress. The axial load reaches a maximum before the material ceases to strain-harden. Thereafter, testing conditions become unstable. The stress corresponding to the maximum is the ultimate, or tensile, strength, which is intended to prevent failure by excessive plastic deformation. Strain-hardening (cold-working) is a cumulative process even if the deformation is reversed. Recovery and recrystallization of the material at almost 0.4 times its absolute melting point removes strain-hardening.

Thixotropy is the tendency of certain substances to flow under external stimuli (e.g., mild vibrations). A more general property is viscoelasticity, a time-dependent transition from elastic to viscous behavior, characterized by a relaxation time. When the transition is confined to small regions within the bulk of a solid, the substance is said to creep. A substance which creeps is one that stretches at a time-dependent rate when subjected to constant stress and temperature. The approximately constant stretching rates at intermediate times are used to characterize the creeping characteristics of the material.

B. Failure of Materials

Materials used in batch processing are subjected to stringent changes of operating conditions during the processing cycle. They fail when (1) they are subjected to stresses beyond the yield point due to accidental runaway; (2) operating conditions become more demanding than those set for design, after several plant expansions; (3) unpredictable conditions due to side reactions or lack of heat dissipation fail to keep intermediates in a thermally stable state; (4) material properties are not as good as expected because of fabrication deficiencies or deterioration by corrosion or embrittlement; (5) sudden changes occur in operating conditions (pressure and thermal cycling and shocking of the materials occur several times per day until the materials fail by fatigue or stress shock).

Failure may be mechanical, due to wear, abrasion and erosion, brittle fracture, surface deterioration, cyclic loading, embrittlement, thermal or pressure shock, or fatigue. Failure may also be chemical, in essence due to corrosion.

Embrittlement is a reduction in the strength of metals caused by hydrogen and caustic substances, probably due to reactions that decarburize steel, thus disintegrating the grain boundaries and promoting the collapse of the crystalline structure. The embrittling effect of hydrogen in steel is reduced by the use of molybdenum and chromium to the extent illustrated in the Nelson diagram.

Corrosion is a reaction, chemical or electrochemical, of the material with the environment. Corrosion resistance often determines the selection of materials for a process. Alloys used in industrial service may require protection and are selected on the basis of the environment and design requirements for each piece of equipment. Corrosion and past service data available from materials of construction and equipment manufacturers are valuable to ensure satisfactory results and long life of plant equipment. Design must have the goal of preventing corrosion in some environments. It should consider the materials and their treatments (liners, coatings, and other alternatives) to minimize trapped moisture, introduction of a new corrosive medium, crevices, and any factors promoting corrosion. Corrosion-resistant materials or methods of protection must be selected for each exposure condition and within prescribed economic limits. Laboratory testing can serve as a guide in this selection, but exposure under actual conditions is necessary in many cases.

Most corrosion is electrochemical, originating with the formation of galvanic cells and the accompanying flow of electrical current. In a metallic medium two dissimilar electrodes may exist because of differences in energy levels, probably due to disordered or stressed areas in the microstructure; differences in composition; or differences in concentration in the electrolytic environment. The electrode with the higher energy potential becomes the anode and suffers corrosion; the cathode is protected. Galvanic corrosion may occur in three different cell types: stress cells, composition cells, and concentration cells. In each, corrosion is produced because one half of a galvanic couple acts as the anode, and the other half, with a lower electrode potential, as the cathode. Only the anode is corroded, when it is in electrical contact with a cathode.

Corrosion can be prevented or reduced significantly by three electrical means.

1. Cathodic Protection

The metal is forced to behave as a cathode; thus, it has no anode areas and does not corrode. This can be achieved in two ways. The first is to apply a large dc current to the corroding metal, which lowers the metal activity to below that of hydrogen; the second is providing for an electrode that acts as an anode. The anode may be inert material, such as graphite or scrap iron, which deteriorates and is

replaced at intervals. To obtain complete protection, the current density must make the anode potential equal to its open circuit potential, at which point no net corrosion can occur. Polarization curves can be used to estimate current density requirements for cathodic protection.

2. Galvanic Protection

This method uses a more active metal than that in the structure to be protected, to supply the current needed to stop corrosion. Metals commonly used to protect iron as sacrificial anodes are magnesium, zinc, aluminum, and their alloys. No current has to be impressed to the system, since this acts as a galvanic pair that generates a current. The protected metal becomes the cathode, and hence it is free of corrosion. Two dissimilar metals in the same environment can lead to accelerated corrosion of the more active metal and protection of the less active one. Galvanic protection is often used in preference to impressed-current technique when the current requirements are low and the electrolyte has relatively low resistivity. It offers an advantage when there is no source of electrical power and when a completely underground system is desired. Probably, it is the most economical method for short life protection.

3. Anodic Protection

Active metals such as aluminum, titanium, and high-chromium steels become corrosion resistant under oxidizing conditions because of a very adherent and impervious surface oxide film that, although one molecule thick, develops on the surface of the metal. This film is stable in a neutral medium, but it dissolves in an acid or alkaline environment. In a few cases, such as certain acid concentrations, metals can be kept passive by applying a carefully controlled potential that favors the formation of the passive surface film. The ability to keep the desired potential over the entire structure is very critical in anodic control. If a higher or lower potential is applied, the metal will corrode at a higher rate, possibly higher than if it is not protected at all.

C. Stress and Fatigue

In batch processing, the reagents and hence the equipment are subjected to cyclic stringent changes in temperature, pressure, and concentration, due to the kind of physical and chemical changes involved. Thermal and pressure stresses arise from the temperature and pressure gradients to which the materials are subjected, since the internal and external layers of metal are subjected to entirely different conditions at the same time. Stresses result even when the gradients are very small, if free expansion or contraction

is prevented by constraints, as in a thermocouple or in a pressure transducer. When the stresses are caused by sudden changes in conditions, the process is referred to as thermal or pressure shock. Stresses caused under shock conditions are greater than those due to slow temperature or pressure changes because of the steeper changes that are generated and the larger rates of application of the stresses. Many materials are affected by the rate at which the load is applied. Some of them are embrittled and unable to withstand a shock stress that they can absorb when it is slowly applied. In dealing with these stresses, it is important to account for plastic flow effects that occur when the yield point is exceeded and also how the flow may change during progressive thermal or pressure cycling of the material. Computational techniques are being developed to account for inelastic effects such as creep and plastic flow, which include cyclic effects in the computational procedure and in the interpretation of material behavior.

Fatigue is a manifestation of a cumulative process leading to progressive fracture under cyclic loading. It starts at cracks of the surface of the material, which propagate inward. It is due to slip concentrated in isolated slip bands inside the grains. It has a statistical behavior in the sense that a population of similar specimens break at widely different numbers of cycles following a normal distribution. Cyclic loading restrains the life of the materials. Plastic strain appears to determine the low-cycle range (<10,000 cycles) of a material. The total strain, elastic plus plastic, is the factor determining the long-cycle range. Alternatively, materials life can be regarded as being governed by stress ranges; this is the total stress to which the material is subjected during the cyclic load. However, the stress range is not as well known as the strain range, and for practical

purposes the cyclic life of materials is expressed as a function of the total strain range. There is a total strain range, ~ 0.006 , below which materials failure does not happen regardless of the number of stress-strain cycles applied. It corresponds to a high-cycle fatigue limit where the life of the material reaches 1 million cycles.

SEE ALSO THE FOLLOWING ARTICLES

CHEMICAL PROCESS DESIGN • FRACTURE AND FATIGUE • PHARMACEUTICALS • REACTORS IN PROCESS ENGINEERING • STOCHASTIC PROCESSES

BIBLIOGRAPHY

- American Institute of Chemical Engineers, Center for Chemical Process Safety, New York.
- Barona, N., and Bacher, S. (1983). "Fundamentals of Batch Processing," Am. Inst. Chem. Eng., New York.
- Chai, C.-P., and Valderrama, J. O. (1982). *Chem. Eng. Sci.* **37** (3), 494.
- Creed, M. J., Fausky, H. K. *et al.* "An easy inexpensive approach to the DIERS procedure," *Chem. Eng. Prog.* **86** (3), 45.
- Knopf, F. C., Okos, M. R., and Reklaitis, G. V. (1982). *Ind. Eng. Chem. Process Des. Dev.* **21**, 79.
- Mauderli, A., and Rippin, D. W. T. (1979). *Comput. Chem. Eng.* **3**, 199.
- Reiner, F., and Musier, H. (1990). "Batch process management," *Chem. Eng. Prog.* **86** (6), 66.
- Renard, M. D. (1979). *Comput. Chem. Eng.* **3**, 9.
- Silver, L. H., Bacher, S., and Hacik, J. (1982). In "Computer Aided Process Plant Design" (M. E. Leesley, ed.), p. 720, Gulf Publishing, Houston, TX.
- Sinha, N. K., and Kuszta, B. (1983). "Modeling and Identification of Dynamic Systems," Van Nostrand-Reinhold, Princeton, NJ.
- Tahamatsu, T., Hashimoto, I., and Hasebe, S. (1982). *Ind. Eng. Chem. Process Des. Dev.* **21**, 431.



Catalysis, Industrial

Bruce E. Leach

University of Texas at Austin

- I. Importance of Industrial Catalysis
- II. History
- III. Mechanisms of Industrial Catalysis Reactions
- IV. Industrial Catalysis Research
- V. Industrial Catalyst Marketing

GLOSSARY

Alkylation Refinery process for the production of high-octane fuel from the reaction of an olefin and isoparaffin in the presence of an acid catalyst.

Applied catalysis Practical or actively used catalysis; catalysis research for a specific purpose.

Attrition Wearing down by friction of a catalyst.

Catalyst Substance that brings about a change in the speed of a reaction without being changed itself.

Catalytic cracking Refining unit to produce distillates from crude oil using an acidic zeolite catalyst.

Catalytic reforming Refinery process by which hydrocarbons of gasoline boiling range are reconstructed with little or no change in carbon number, leading to an improvement in fuel quality; reactions in reforming include isomerization, hydrogenation, dehydrogenation, and dehydrocyclization.

Fundamental catalysis Work that develops the basic principles and laws of the science of catalysis.

Hydrocracking Refinery process to convert high-boiling molecules to lower boiling molecules by hydrogenation and carbon bond breaking.

Hydrotreating Refining process to remove sulfur, nitrogen, and oxygen from petroleum feedstocks by contacting with hydrogen in the presence of a nickel/molybdenum or cobalt/molybdenum on alumina catalyst.

Oxychlorination Process to produce vinyl chloride monomer from ethylene, hydrogen chloride, and oxygen over a copper chloride on alumina catalyst.

Proprietary Information and technology that is not public knowledge and is to be kept confidential.

Reclamation Renewal or restoration of a catalyst to initial chemical and physical properties.

Selectivity Percentage of a desired reaction product compared with the theoretical quantity of that reaction product.

Stream factor Percentage of actual plant on-stream time compared with the possible on-stream time.

INDUSTRIAL CATALYSIS is the commercial process of finding and enhancing the performance of substances that increase the rate at which a chemical reaction reaches equilibrium. Industrial catalysis is vitally concerned with the activity, selectivity, lifetime, and environmental impact of

a catalyst. It is highly competitive and must be profitable. The scope of industrial catalysis includes catalyst invention, theory, development, physical properties, poisoning, replacement schedules and techniques, regeneration, disposal, and competitive trends.

Industrial catalysts are commonly divided into petroleum, chemical, polymerization, and environmental catalyst segments of the market. Companies seek to leverage their expertise in surface chemistry and materials science to develop products and systems that, in turn, improve customers' products.

New and improved catalysts are central to the competitiveness of individual products, companies, industries, and countries. Improved catalysts are often the lowest investment cost route to feedstock and energy savings. Catalysts reduce costs by increasing the selectivity for a product and/or allowing a process to operate at lower temperature or pressure. The catalyst industry has recently undergone globalization, consolidation, restructuring, and downsizing which affects catalyst development and marketing. Industrial catalysis is truly technology-based and market-driven.

I. IMPORTANCE OF INDUSTRIAL CATALYSIS

A. Economics

Catalyst sales are a \$10 billion per year industry. Catalysts effect on the economy is much larger if the value of the products formed from catalyzed reactions is also counted. An estimated 90% of plastics, chemicals, pharmaceuticals, fuels, and other consumer products are made using catalysts (Table I).

The catalyst industry has grown from its beginnings in the early 1900s at a rate paralleling the increase in the United States gross national product. Catalysis has been a major factor in decisions to design, construct, and operate plants to produce new products. In the United States over 7000 different products worth an estimated \$375 billion per year are produced via industrial catalysis.

Employment in industrial catalysis is primarily at catalyst, chemical, and oil companies. However, catalysis is a highly diversified field and a wide variety of professionals are employed by the industry. The catalysts themselves are specialty products requiring the application of science, engineering, and proprietary art.

Major innovations in the chemical and petroleum industries have involved breakthroughs in catalyst technology. A new or improved catalyst is often the basis for a competitive manufacturing cost advantage. Sources of cost savings

TABLE I Common Industrial Catalysts

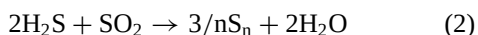
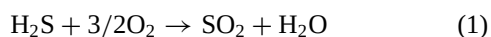
<i>Refinery catalysts</i>	
Catalytic cracking	Zeolites, silica-alumina
Reforming	Pt/Re on alumina
Alkylation	HF or H ₂ SO ₄
Hydrotreating	NiO/MoO ₃ ; CoO/MoO ₃ on alumina
Hydrocracking	Pt on zeolite or alumina
Hydrosulfurization	NiO/MoO ₃ ; CoO/MoO ₃ on alumina
<i>Environmental catalysts</i>	
Automobile emissions	Pt/Pd/Rh on alumina
Hydrocarbon emissions	Pt/Pd on alumina
Claus	Alumina
<i>Polymerization</i>	
Polyethylene (Phillips)	CrO ₃ on silica
Polyethylene (Ziegler-Natta)	AlR ₃ , TiCl ₃ , MgCl ₂ + other, Metallocene
Polypropylene	Metallocene, Ziegler-Natta
Polyvinyl chloride	Organic peroxides
Polystyrene	Benzoyl peroxide
Polybutadiene	AlR ₃ , TiCl ₃
Polyformaldehyde	BF ₃
Polyphenylene oxide	Cu(I) or Mn(II) with amines
Polycarbonates	AlCl ₃
<i>Chemical catalysts</i>	
Acetylene hydrogenation	Ni on alumina (sulfided) Pd on alumina (CO-poisoned)
Acrylonitrile	Noble metals/C or Ni
Alcohols from esters	CuCr ₂ O ₄
Amines from nitriles	Raney Ni
Ammonia	Fe ₃ O ₄ /Al ₂ O ₃ /K ₂ O melt
Butene from butane	NiO/Al ₂ O ₃
Carbon monoxide shift (high temp.)	Fe ₂ O ₃ /Cr ₂ O ₃
Carbon monoxide shift (low temp.)	CuO/ZnO
Desulfurization	ZnO
Ethylene oxide	Ag on support
Formaldehyde	Fe(MoO ₄) ₃
Hydrogenated oils	Ni in oil
Maleic anhydride	V ₂ O ₅
Methanation	Ni or Ru on alumina
Methanol (high pressure)	ZnCr ₂ O ₄
Methanol (low pressure)	Cu-chrome-Zn
Nitric acid	Pt/Rh gauze
Oxychlorination of ethylene	CuCl ₂ /KCl/Al ₂ O ₃
Paraffin dehydrogenation	Pt/Al ₂ O ₃ (S-poisoned)
Phthalic anhydride	V ₂ O ₅
Steam reforming	NiO/AlO
Sulfuric acid	V ₂ O ₅ /K ₂ O/SiO ₂

due to catalyst changes include: (1) reduced feedstock consumption or cost; (2) lower energy consumption; (3) increased by-product credits or reduced by-product debits; (4) reduced capital investment; and (5) increased stream factor.

Process improvements can also result in products of higher quality or in a safer, more environmentally acceptable commercial operation. Catalysts of higher activity can reduce the required operating temperature or pressure in a process unit and save energy.

B. Environment

Catalysts help customers comply cost-effectively with clean-air regulations. Hydrocarbons, carbon monoxide, and nitrogen oxides can be removed using supported precious metal catalysts. Organic sulfur compounds are converted to H_2S using nickel/molybdenum or cobalt/molybdenum on alumina catalysts. Sulfur can be recovered in a Claus process unit. The Claus catalytic converter is the heart of a sulfur recovery plant.



The first reaction takes place at high temperature in a furnace fed with a sour gas and air mixture. The second reaction is catalyzed by alumina. Operational temperatures are $\sim 330^\circ\text{C}$.

The most widely known pollution control catalysts are those for auto emission control. Automotive catalysts can be of two types—monoliths and pellets. Monoliths now dominate the market. Pollution control catalysts are also used to control diesel emissions.

Automotive emission control is a major catalyst market segment. These catalysts perform three functions: (1) oxidize carbon monoxide to carbon dioxide; (2) oxidize hydrocarbons to carbon dioxide and water; and (3) reduce nitrogen oxides to nitrogen. The oxidation reactions use platinum and palladium as the active metal. Rhodium is the metal of choice for the reduction reaction. These three-way catalysts meet the current standards of 0.41 g hydrocarbon per mile, 3.4 g carbon monoxide per mile, and 0.4 g nitrogen oxides per mile.

New proposed “Tier 2” emission standards proposed for introduction in the 2004 model year would apply to sport utility vehicles, minivans, and pickup trucks and make them meet the same standards as passenger cars. It has also been proposed to lower the sulfur level in gasoline to 30 PPM from the current average level of 300 PPM by 2004. The Euro IV limits proposed for 2005 are satisfied by technology from at least one catalyst supplier

already. Implementation of these proposals will increase environmental catalyst markets.

C. Standard of Living

Industrial catalysts have made it possible to utilize and enjoy many new products in the areas of plastics, transportation, clothing, detergents, food supply, and construction. The production of most polymers involves catalysis either in polymerization or in monomer synthesis. Improved fuels, tires, and construction materials have revolutionized the transportation industry in this century. Synthetic fibers are widely used in clothing and carpets. Biodegradable detergents are available for inexpensive cleaning. Fertilizers, pesticides, and herbicides have been used to increase crop yields to feed a growing world population.

II. HISTORY

Catalysis has made possible the change in the chemical process industry from feedstocks of coal and acetylene, to ethylene. Activation of alkanes is now a major research topic. German industrial scientists led in the coal- and acetylene-based chemical industry developments. Many of the chemical products were for the dyestuffs industry.

Multicomponent catalysts were first studied after 1900 at Badische Anilin-und Soda-Fabrik (BASF). This led to the discovery of a magnetite promoted with alumina and alkali ammonia synthesis catalyst in 1908 by Haber. Bosch and BASF developed a methanol synthesis catalyst composed of the mixed oxides of zinc, chromium, and potassium in 1923. Fischer and Tropsch made synthetic hydrocarbons from synthesis gas in 1927. In the United States, Union Carbide made chemicals initially from acetylene. They became interested in ethylene and in 1926 began production of cellosolve based on ethylene. In 1927 Shell Chemical began steps to produce chemicals from petroleum feedstocks. Initially they focused on ammonia, propylene, and solvents. Standard Oil Company (NJ) about the same time decided to study the application of chemical engineering to the upgrading of petroleum fractions. This led them into hydrogenation, olefins, and aromatics catalysis research. Dow Chemical led in the development of chlorine and bromine chemistry using natural gas as the energy source.

There is a continuing stream of new-generation catalysts for refining, polyolefin formation, oxychlorination, hydrogenation, and other catalyst applications. Some of the names and areas of contribution in industrial catalysis are given in [Table II](#). Highlights of industrial application of catalysts are given in [Table III](#).

TABLE II Scientists and their Contributions to Industrial Catalysts

Catalysis scientist	Company	Contribution
Eugene Houdry	Sun Oil, Mobil Oil	Hydrocracking
Herman Pines	UOP, Inc.	Catalytic cracking
Irving Langmuir	General Electric	Adsorption theory
Vladimir Ipatieff	UOP	Catalytic cracking
Vladimir Haensel	UOP	Platforming process
Paul Emmett	Fixed Nitrogen Laboratory	Ammonia synthesis
Fischer and Tropsch	Ruhrchemie	Fischer-Tropsch
Haber and Bosch	BASF	Ammonia synthesis
Otto Beeck	Shell Oil	High vac analysis films
Ernest Thiele	Standard Oil (Indiana)	Role of diffusion
Charles Plank	Mobil	Zeolite cracking
Frank Ciapetta	W. R. Grace	Refinery catalysts
Thomas Singleton, et. al.	Monsanto	Methanol carbonylation
Paul Hogan, Robert Banks	Phillips Petroleum	Polyolefin catalysts
Robert Banks	Phillips Petroleum	Olefin disproportionation
A. Mittasch	BASF	Methanol synthesis
Leonard Drake	Mobil	Mercury porosimetry
Cambell, Jahnig, Martin	Exxon	Fluid catalytic cracking
Robert Grasselli	Standard Oil (Ohio)	Ammoxidation
R. Eischens	Texaco	Infrared adsorbed CO
Scott and Sullivan	Chevron	Isocracking
Karl Ziegler,	Max Planck Institute	High density polyethylene
Waldo Semon	BF Goodrich	PVC

III. MECHANISMS OF INDUSTRIAL CATALYSIS REACTIONS

Mechanisms of reactions are important for industry because they provide information useful for optimizing catalyst and reactor conditions. The study of reaction mechanisms in industry cannot stand alone as it can in academia. Mechanistic studies are funded to solve plant problems, to decrease operating costs, and to improve product quality. There is a wide variation in industry in the amount and type of mechanistic research funded and the timing for such research. Mechanistic research on chemical reactions is most easily justified when it is focused on the development of commercial products for a company. Often the results of mechanistic studies are not published but used instead in reactor modeling. The second reason is that competitors would obtain the information at no cost.

TABLE III Industrial Catalysis Advances

First plant approximate date	Development	Company
1906	Nitric acid	Hoechst
1913	Ammonia	BASF
1921	Tetraethyl lead antiknock	GM
1922	Production TEL	Standard (New Jersey)
1928	Diethylene glycol	Union Carbide
1930	Synthetic ethanol	Union Carbide
1931	Ammonia from natural gas	Shell Oil
1931	Methyl ethyl ketone	Shell Oil
1931	Fixed bed catalytic cracking	Mobil, Sun
1934	Fischer-Tropsch	Ruhrchemie
1934	Bromine	Dow
1935	Isopropanol and acetone	Shell Oil
1937	Styrene	Dow
1937	Nylon	DuPont
1939	Alkylation of paraffins	
1942	Fluid catalytic cracking	Esso
1942	Butadiene	Shell
1943	Butadiene by dehydrogenation	Humble
1964	Molex paraffin separation	UOP
1970	Methanol— low pressure	ICI
1971	Parex <i>p</i> -xylene separation	UOP
1973	Acetic acid— low pressure	Monsanto
1973	Xylene isomerization	Mobil
1974	Kevlar	DuPont
1978	Ethylene glycol via acetoxylation	Halcon
1985	Methanol to gasoline	Mobil
1995	Detal solid acid alkylation	UOP, Petresa

Mechanistic studies start with determination of the kinetic rate law and the rate-limiting step; information on heat and mass transfer is also needed. These studies may use such techniques as isotopic labeling, chemisorption measurements, surface spectroscopy, temperature-programmed desorption, and kinetic modeling experiments.

The design of a catalyst requires knowledge of the reaction mechanism to modify the catalyst surface sites

intelligently, thereby increasing the rate of the desired reaction and minimizing the side reactions and their products. Surface area, pore volume, pore volume distribution, acidity or basicity, the number of active sites, and the chemical surroundings of these active sites must be fixed for a catalyst. Reactants must be adsorbed on the active site, chemical interactions must occur that are neither too strong or too weak, the transition state intermediate must react with other molecules or rearrange, and finally the reaction products must diffuse away from the reaction site. Catalysts are often designed with optimal support geometry with active metals deposited at an optimum distance from the center for maximum selectivity.

Analytical instrumentation is making studies of mechanisms of industrial reactions under operating conditions more feasible. A number of mechanisms of important commercial reactions are still not known conclusively. Sometimes it is almost impossible to distinguish among several alternatives. Refinements or sometimes even complete changes in mechanisms have been made as analytical capabilities have improved.

Often model systems are used in mechanistic studies. Use of a single-component feed rather than a broad boiling range greatly simplifies analysis schemes. However, part-per-million quantities of impurities in real feedstocks can sometimes severely complicate catalyst usage, selectivity, and life.

Examples will be taken from the reactions listed in Table I to illustrate mechanisms and their effect on commercial catalyst development.

1. *Catalytic cracking.* Catalytic cracking of hydrocarbons to produce lower molecular weight hydrocarbons occurs by the heterolytic cleavage of a C—C bond. A carbonium ion mechanism is involved. Strong acids such as silica-alumina or zeolites are used as commercial catalysts. Hydrogen transfer activity relative to β -carbon scission is important in selectivity. Alkane and aromatic products are preferable to olefins. Highly olefinic products coke a catalyst faster and neutralize acid sites. Catalysts can be regenerated by controlled thermal oxidation in air.

2. *Reforming.* Reforming is used to upgrade the octane number of gasoline. In the presence of hydrogen a desirable catalyst should promote isomerization, cyclization, and arenes. Such a catalyst has a dual function; it is acidic to promote skeletal rearrangement by carbonium ion mechanism, and it has a hydrogenation-dehydrogenation component to promote arene and cyclization reactions.

3. *Alkylation.* Reactions of olefins with isoparaffins require a highly acidic catalyst. Both Friedel-Crafts and protonic acids are used. The protonic acids, sulfuric (96–100%) and hydrofluoric acids, are commonly used. If the

acid strength is too low, olefin polymerization becomes a serious side reaction. Considerable effort is in progress to use zeolites and solid acid catalysts to replace liquid acids in alkylation.

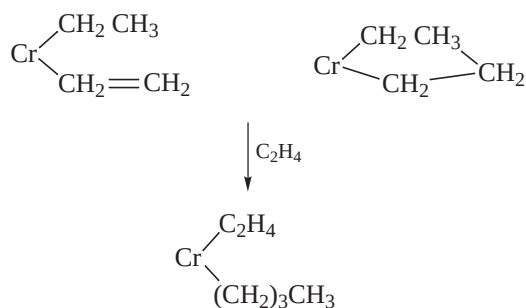
4. *Hydrotreating and hydrodesulfurization.* The major reactions are desulfurization, denitrogenation, and olefin saturation of petroleum feedstocks. Aromatic saturation is generally undesirable, so hydrogenation activity must be moderate. The nickel/molybdenum on alumina catalyst must be sulfided to achieve desired activity and selectivity. Sulfur is removed as hydrogen sulfide; nitrogen is removed as ammonia. Substrate acidity should be moderate to low or excessive coke is formed. Diffusion is an important factor, especially for heavy crudes.

5. *Automobile and Hydrocarbon Emissions.* The oxidation of carbon monoxide and hydrocarbons is catalyzed by platinum/palladium/rhodium on alumina. If catalyst poisons such as lead and phosphorus are not present, the major problems become initiation of oxidation at low temperature, thermal stability at high temperature, resistance to thermal shock, and a high external surface area catalyst configuration.

If nitrogen oxide control is one of the catalytic requirements, the stoichiometry of air-to-fuel ratio must be kept nearly stoichiometric to reduce NO; then air must be added and CO and hydrocarbons oxidized in a second part of the catalyst bed.

The vapor-phase, high-space-velocity oxidation to the thermodynamic reaction products should be contrasted with kinetically controlled oxidation of chemical feedstocks when the active metal is purposely poisoned or the surface area reduced.

6. *Polyethylene (chromium catalyst).* The chromium on silica catalyst is quickly reduced from Cr(VI) to Cr(II). The active site consists of a single chromium ion present as silyl chromate before reduction with ethylene. Ethylene adds to the chromium as indicated.



Termination of the growing chain can occur by hydride transfer to the active site instead of to the monomer.

7. *Polyethylene (Ziegler-Natta catalyst).* Most commercial catalysts start with titanium tetrachloride, diethylaluminum chloride, and magnesium chloride as a

support, as well as other promoters. Stereoselectivity (for polypropylene and higher olefins) is controlled by addition of aromatic esters such as ethyl benzoate.

8. *Polyethylene and polypropylene (metallocenes).* Metallocenes of different types are being used in a variety of commercial processes to make polymers with different properties than traditional Ziegler-Natta catalysts. The metallocene catalysts can be optimized for chain length and stereochemical control of the product polymer.

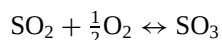
9. *Polyvinyl chloride.* Organic peroxides are used to catalyze the free radical polymerization of vinyl chloride monomer in water. The organic peroxide is selected to generate free radicals thermally at the temperature of polymerization.

10. *Polystyrene.* The polymerization of styrene is most commonly done under free radical conditions. Peroxides are used to initiate the reaction at low temperatures. At $\sim 100^\circ\text{C}$ styrene acts as its own initiator. Below 80°C the termination mechanism primarily involves combination of radicals. Above 80°C both disproportionation and chain transfer with the Diels-Alder dimer are important.

11. *Polybutadiene.* Most polybutadiene is made by an emulsion process with a free radical initiator. If stereoregular cis-1,4-polybutadiene is desired, a titanium-based Ziegler-Natta catalyst is used. The catalyst is similar to those used for polyethylene and polypropylene in type and mechanism.

12. *Polyformaldehyde.* Polyformaldehyde or polyacetal is made by two different processes. Delrin is made from formaldehyde by anionic polymerization catalyzed by a tertiary amine. The homopolymer is end-capped with acetic anhydride. Celcon is made from trioxane cationic copolymerization using boron trifluoride catalyst and ethylene oxide (2–3%) as the comonomer. Boron trifluoride is a Lewis acid that associates with trioxane and opens up the six-membered ring. Ethylene oxide provides the end capping. Without an end cap, polyformaldehyde is thermally unstable and loses formaldehyde units.

13. *Sulfuric acid.* The oxidation of SO_2 to SO_3 is the step in sulfuric acid manufacture that requires catalysis. The oxidation is exothermic, and the equilibrium becomes more unfavorable for SO_3 at higher temperatures.



The rate law can be expressed as

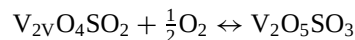
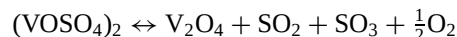
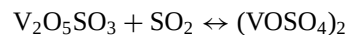
$$R_f = k p^x(\text{SO}_2) p^y(\text{O}_2) p^z(\text{SO}_3)$$

Where x and y are between 0.5 and 1.0, and z is usually 0.0 to -1.0 .

Normal operating temperature for the vanadium catalyst is $450\text{--}550^\circ\text{C}$.

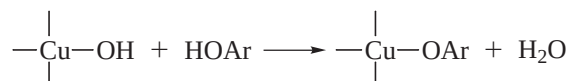
Multiple stages are used for heat transfer and to prevent temperatures above 600°C , which can damage the catalyst.

A mechanism for catalyst oxidation-reduction (others can be written) is the following:



14. *Polyphenylene oxide.* Oxidative polymerization of 2,6-xenol to the engineering resin polyphenylene oxide (PPO) is catalyzed by copper and manganese amines. Pyridine is a typical amine used in the polymerization.

The active copper catalyst is $\text{Cu}(\text{Cl})(\text{OH})(\text{NR}_3)_2$. The first step in the reaction is the following, where ArOH stands for 2,6-xenol:



Electron transfer from oxygen to copper gives a phenoxyl radical, which couples with another copper-bound radical to form the $\text{C}-\text{O}-\text{C}$ dimer and $\text{Cu}(\text{I})$. The reaction behaves as a step reaction rather than a chain reaction. A quinol ether rearrangement occurs to equilibrate polymer and monomer. High-molecular-weight polymer is formed only in the late stages of reaction. Indeed, other phenols are incorporated into the polymer if they are added at the end of the reaction because of the quinol ether rearrangement.

15. *Polycarbonates.* Phenol and phosgene react under basic (sodium hydroxide) conditions to form diphenyl carbonate. Bisphenol A and diphenyl carbonate are melted together with a small quantity of basic catalyst (Na , K , Li) $_2\text{CO}_3$. The temperature is slowly raised to 250°C , and phenol is removed in the polymerization step.

16. *Acetylene hydrogenation.* Selective hydrogenation of acetylene to ethylene is performed at $\sim 200^\circ\text{C}$ over sulfided nickel catalysts or carbon-monoxide-poisoned palladium on alumina catalyst. Without the correct amount of poisoning, ethane would be the product. Continuous feed of sulfur or carbon monoxide must occur or too much hydrogen is chemisorbed on the catalyst surface. Complex control systems analyze the amount of acetylene in an ethylene cracker effluent and automatically adjust the poisoning level to prepare the catalyst surface for removing various quantities of acetylene with maximum selectivity.

17. *Alcohols from esters.* The major problem is reaction selectivity. Paraffin by-product in alcohol results if the catalyst activity is too high. Yet the reduction of esters to alcohols is a difficult reaction. Copper chromite catalyst, 3000–5000 psig hydrogen, and a temperature of $270\text{--}300^\circ\text{C}$ are required for the reduction. An alternate catalyst is CuO/ZnO , which is used for methyl ester reduction only. Hydrogen solubility in alcohol is limiting.

The rate law for the hydrogenation of esters to alcohols shows a dependence on the square of the hydrogen pressure.

18. *Ammonia*. The rate law for ammonia synthesis,

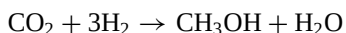
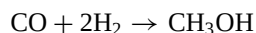
$$\text{Rate} = k_1 + p\text{N}_2 x(p^3\text{H}_2/p^2\text{NH}_3)^a - k_2(p^2\text{NH}_3/p^3\text{H}_2)^{1-a}$$

takes the forward and reverse reactions into account. Potassium oxide, alumina, and calcium oxide are promoters for iron, which is the basic catalytic material. The promoters increase basicity, stabilize surface area, and increase reaction rate. Chemisorption of nitrogen (dissociative) is the rate-limiting step. The product ammonia competes for active sites, so the reaction is run at relatively low conversion. Catalysts have been improved, allowing operations at lower pressure and longer catalyst lifetimes.

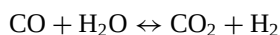
19. *Ethylene oxide*. The oxidation of ethylene to ethylene oxide is exothermic (~ 117 kJ/molH), and further oxidation to carbon dioxide and water is even more favorable thermodynamically. The reaction must be run under kinetic control to prevent total oxidation.

Oxygen and ethylene adsorb on a supported silver catalyst. A Rideal-Eley mechanism is favored for the adsorption. Chloride is a promoter/modifier for the catalyst and reduces the surface oxygen content. Silver(I) on the surface tends to increase ethylene adsorption, which explains why small levels of chloride increase the rate of reaction.

20. *Methanol*. Methanol can be synthesized from mixtures of carbon monoxide, carbon dioxide and hydrogen:



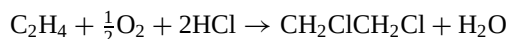
The shift reaction equilibrium is also important:



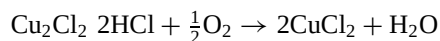
The rate expression can be simplified if the assumption is made that methanol desorption is the rate-limiting step on a Cu/ZnO/Al₂O₃ catalyst:

$$\text{Rate} = k(p^2\text{H}_2)(p\text{CO})^{0.7}$$

21. *Oxychlorination of ethylene*. Knowledge of the role of copper chloride and the mechanism of oxychlorination has evolved. Current theory suggests that Cu(II) chloride chlorinates ethylene, which is chemisorbed on the catalyst.



Ethylene chemisorption should be enhanced at Cu(I) sites, which result from the chlorination of ethylene. Cu₂Cl₂ is reoxidized by HCl and oxygen to CuCl₂:



Oxychlorination of ethylene is highly exothermic, and heat removal from reactors must be efficient. Graded-activity catalysts are used in fixed-bed reactors, or fluid-bed reactors are used to control heat generation and transfer. Inert substances can be used to moderate conversion in a bed, or the KCl loading on the catalyst can be varied to control reaction rates. Advances have been made using oxygen instead of air. Catalyst life has been improved significantly by changing the catalyst shape (rings and spheres) to decrease carbon formation in the interior of the alumina catalyst matrix.

IV. INDUSTRIAL CATALYSIS RESEARCH

A. Research For Profit

Industrial catalysis research can be both fundamental and applied. The long-term objective must be to make a profit using the knowledge of catalysis acquired through research. Different project evaluation schemes are used to assess the potential profit of catalysis-related research. Such factors as company size, business philosophy, competition, and economic conditions affect the ratio of applied to fundamental catalysis research.

The high cost of industrial research has prompted industry-government, inter-industry, and industry-university alliances to fill the research void.

Examples of industry-government alliances are the National Institute of Standards and Technology Advanced Technology Program, the Japanese Agency of Industrial Science and Technology which operates 15 laboratories, the National Center for Scientific Research in France, and the Italian National Agency for New Technology.

Industry-university alliances are illustrated by Centers for Catalytic Science and Technology such as those at the University of Delaware and Delft Department of Chemical Technology.

Illustrative of intercompany alliances in catalysis is the hydroprocessing MAKFinning Technologies. This combines technologies and expertise from Mobil, Akzo Nobel, M. W. Kellogg, and Fina. The goals of the alliance include, "to create innovative approaches and optimum solutions to the industry's changing product demands and to make a more effective use of each company's research and development resources."

Another way to leverage the scarce research and development funds is to hire a catalyst-consulting group. Several catalyst-consulting companies exist to work on a confidential basis on projects sponsored by clients with an interest in business development, market strategy analysis, technological benchmarking, and new product and process innovation.

B. Finding a Competitive Advantage

Technical information is a valuable asset to a company. A patent gives the owner the right to prevent others from making, using, or selling an invention for a period of time. Patents often describe cost savings in processing that can be translated into greater profit margins or a new profitable product. Patents can also be licensed or sold as a source of revenue.

As an example, 28% of the world's gasoline owes its existence to an idea to use synthetic zeolites to make more gasoline from crude oil. It has saved consumers billions of dollars a year and it has greatly extended oil reserves.

Some information is vitally important in the operation of a business, but either it cannot be patented, or if patented the patent would be difficult to enforce. Such information is held as a "trade secret." Catalyst preparation techniques are good examples of proprietary information that a company may choose not to patent.

Catalyst performance must continually improve to keep a product technically superior. Successful catalyst research programs coupled with production cost minimization may allow a plant to be run at capacity while its competitor shuts down or operates at a lower capacity and reduced profit margin. In many instances, finding the competitive advantage is essential to business survival. Finding or making the most cost-effective catalyst is a strong incentive for catalysis research.

C. Solving Plant Problems

Catalyst performance is monitored in a commercial unit from start-up to change-out. Typically this is done using statistically designed experimentation. This results in projections of new catalyst demand and information on how process variables affect catalyst performance.

Companies seek to make their plants more efficient and cost competitive. New catalysts are screened, alternative feedstocks tested, and changes in process equipment evaluated to improve product quality and production economics. The risk of making changes in a plant without laboratory or pilot plant work is usually too great, although here are exceptions.

Bench scale or pilot plant research is valuable to determine corrective measures to take when unexpected circumstances occur. Operational mistakes made in plants and resulting in down-time are expensive. Operational problems can arise because of feedstock changes, instrument errors, leaking valves, etc. It is useful to know the effect of all potential poisons in a process and corrective measures to take in unit operation.

Future process development is another large area of catalysis research. This can entail the search for new prod-

ucts or accommodating major feedstock, catalyst, or reactor changes. Major areas of future interest are environmental protection and the conversion of paraffin's and agricultural commodity feedstocks to industrial chemicals.

D. Basic Catalysis

Some industrial catalysis research is theoretical and some is at the frontiers of science. Often a process is selected for commercialization before a commitment is made to the basic research. The open technical literature is not a good indication of the quantity or quality of basic catalyst research work done in industry. Many of the results of industrial catalysis research are patented or kept secret. Publications usually follow patents and may represent research completed 3–5 years previously or work on a project that was not commercialized for economic or technical reasons that are not obvious.

As noted, industry is partnering with government for high risk and enabling technologies. Four particular areas of interest were identified by industry for future research:

1. Major yield and selectivity improvements to reduce waste and energy consumption, minimize feedstock costs, or enable market entry of new feedstocks
2. Clearer structure/function relationships to better predict and/or control catalyst structures linked to performance metrics, and reduce the time to market for higher performance products and processes at lower cost
3. New catalyst uses and fabrication methods to minimize emission abatement costs
4. Innovative reactor configurations that enable better integration of transport processes with catalytic performance for reduced capital and operating costs

A growing area of research is biocatalysis. The challenge is to use renewable agricultural products as feedstock and convert these feedstocks cleanly and selectively to chemicals using enzymes. Catalysis, bioengineering, and chemical engineering will be involved in development of new technologies.

Catalysis clubs have been organized in a number of countries. The internet is a good source of current information, for example, the North American Catalysis Society web page can be located at www.dupont.com/nacs. This web site contains a link to the International Congress on Catalysis, which meets every four years. Awards and local meetings for catalysis scientists are also advertised on the internet web page.

V. INDUSTRIAL CATALYST MARKETING

Many companies that do industrial research on catalysis choose not to make their own catalysts. Catalyst preparation and marketing is a specialty chemical, high technical service business. Manufacturers are under pressure to make their catalysts more active, more selective, and with a greater cost performance and lifetime than those of their competitors. Development of a new catalyst or process historically has taken many years (5–10) which is a disadvantage in project economics. The overall catalyst business is expanding and catalyst life is finite. The challenge is to make a cost-effective product with a sufficiently high rate of return on investment.

Catalysts are usually subdivided into homogeneous and heterogeneous classes. Homogeneous catalysts are soluble in the reaction media. Heterogeneous catalysts make up the bulk of the catalyst market; they are the solid catalysts that can be a support material, such as alumina, silica, or silica alumina, but more often some metal salt is added to a formed catalyst support.

Petroleum catalysts constitute the largest catalyst market, followed by automobile emissions catalysts and chemical catalysts. Catalytic cracking is the largest volume petroleum application, followed by alkylation, hydrotreating, hydrocracking, and catalytic reforming. The highest volume chemical catalysts in order are polymerization catalysts, oxidation catalysts, ammoxidation, oxychlorination, ammonia synthesis, methanol synthesis, hydrogenation, and dehydrogenation.

A. Scale-Up and Development

Catalyst scale-up is a process in which a catalyst previously made in small quantities in a laboratory is manufactured in quantities of more than 100 lb with equipment that performs the same operations as larger commercial equipment. If possible, this should involve equipment that simulates the series of unit operations that will be used in commercial catalyst preparation.

In this scale-up stage an objective is to simplify the laboratory preparation while maintaining the essential activity and selectivity of the catalyst. Unit operations available at most custom catalyst manufacturers include precipitation, gel formation, filtration, washing, impregnation, coating, kneading, drying, calcination, grinding and sieving, dry-mixing, tableting, extrusion, beading, leaching, melting, and activation. If special treatment of the catalyst is needed, it may add considerable expense to the catalyst preparation.

A catalyst made in the laboratory or by a catalyst manufacturer must be tested and meet quality control specifications to ensure high quality and reproducibility. Catalyst companies are ISO 9002 certified which also provides

TABLE IV Catalyst Quality Tests

Activity	Laboratory activity test
Surface area	Gas adsorption
Pore volume	Water adsorption
Pore volume distribution	Mercury porosimetry
Metal surface area	Selective gas adsorption
Thermal stability	Thermal gravimetric analysis
Crush strength	Compression test
Attrition resistance	Rotary drum test
Dimensions	Measurement
Particle size	Sieve analysis
Density	Weight per volume
Purity	Trace metal analysis
Metal distribution	Electron microprobe
Lifetime	Accelerated lab test
Crystalline form	X-ray
Crystallite size	X-ray

quality assurance. Common catalyst test procedures are outlined in Table IV. Quality control testing by catalyst vendors and catalyst purchasers is vital for minimizing potential commercial problems. This is most important to the vendor, who develops a reputation (or lack thereof) for excellence in product quality.

B. Proprietary Technology

Proprietary information is often transferred between companies after both sign a secrecy or confidentiality agreement. Generally, the agreement defines what information is to be exchanged, limits how long it is to be held confidential (2–10 years is common), prohibits third-party disclosures, and provides exceptions. Common exceptions are if the information transferred is already known or if the information becomes public at a later time.

Custom-catalyst manufacturers build a reputation based on keeping proprietary information confidential, and their overall performance in this area is excellent. The legal departments of the companies involved agree to the terminology of the agreement, and the agreement is signed by a vice president or other senior company official. These agreements permit joint projects to be undertaken between catalyst inventors and catalyst suppliers in which catalyst recipes are transferred and optimized.

C. Licensing

Catalyst formulations and technology can be licensed from inventors. The legal department of the company offering the technology draws up the license. It can be exclusive or nonexclusive. For example, hydrocracking catalysts are usually licensed along with a process, whereas

fluid catalytic cracking catalysts are items of commerce. License agreements can be quite complicated legal documents. Some affect ownership of catalyst developments in the future, and process guarantees stipulate the catalysts to be used.

Licensing is a source of income and also controls the extent of technology transfer. The purchaser of technology is restricted from passing information to third parties about the catalyst or process in most license agreements.

D. Service to Customers

Catalyst manufacturers are continually seeking new catalyst markets. Small catalyst samples for testing are generally available at no or moderate cost. Sometimes these require confidentiality agreements as discussed earlier or a nonanalysis agreement. This pledges the recipient of the catalyst not to try to determine the catalyst composition or method of preparation. If catalyst performance in a process is the objective, these restrictions are no problem.

Larger pilot-scale catalyst samples are often contracted on a per day charge to the potential customer, or a bid is made on the preparation of a fixed quantity of final product that meets agreed-on specifications.

Catalyst companies may provide catalyst testing support. This support is to maintain and develop new markets for their products.

When a catalyst is sold, the catalyst vendor often provides technical support to ensure that the catalyst is properly loaded and any pretreatment steps are correctly completed. This service minimizes the chances of problems developing later with the commercial unit. If problems develop with the catalyst during its expected lifetime, the catalyst marketing group again provides technical support in the form of catalyst testing. This is expected within the industry and is further evidence of the intense competition in catalyst marketing. An excellent sales and technical service organization is essential for the continuing success of any commercial catalyst manufacturer.

E. Reclamation and Regeneration

Catalysts deactivate during use. The most common reason for deactivation is coke or carbon formation on and in the catalyst particles. The carbon can be burned off under carefully controlled conditions to regenerate the catalyst with activity and selectivity very nearly like that of a new catalyst. In the removal of carbon by oxidation, the quantity of oxygen and the temperature to which the catalyst particles are exposed must be limited.

Some refinery catalysts are regenerated repeatedly as part of the commercial process (e.g., FCC catalysts). In this case the regeneration facility is part of the on-site process. Regeneration can be performed with rental equip-

ment on site, but a growing industry is commercial off-site catalyst regeneration. Spent catalyst is removed and transferred to the regeneration firm's plant, where the catalyst is regenerated, screened, and sent back for a toll fee. This provides a number of advantages to the catalyst user, including the following. The investment in regeneration equipment is reduced, there is less technology to master, and the custom regenerator meets the environmental standards.

Controlled burning of carbon does not regenerate all catalysts. Catalysts can be deactivated by particle growth, compound formation, tramp metal deposition, crystal-phase changes, and adsorption of catalyst poisons that cannot be reversed by thermal oxidative treatment.

SEE ALSO THE FOLLOWING ARTICLES

ADSORPTION (CHEMICAL ENGINEERING) • CATALYST CHARACTERIZATION • CATALYSIS, HOMOGENEOUS • INCLUSION (CLATHRATE) COMPOUNDS • KINETICS (CHEMISTRY) • PETROLEUM REFINING • PHARMACEUTICALS

BIBLIOGRAPHY

- Armor, J. N., ed. (1994). "Environmental Catalysis (ACS Symposium, No. 552)," *Am. Chem. Soc.*, Washington DC.
- Davis, B., and Hettinger, W., Jr., eds. (1983). "Heterogeneous Catalysis: Selected American Histories," *Am. Chem. Soc.*, Washington DC.
- Gates, B. C. (1991). "Catalytic Chemistry (The Wiley Series in Chemical Engineering)," Wiley, New York.
- Hegedus, L. L. (1987). "Catalyst Design: Progress and Perspectives," Wiley, New York.
- Herkes, F. E., ed. (1998). "Catalysis of Organic Reactions (Chemical Industries Series, vol. 75)," Marcel Dekker, New York.
- Imelik, B., and Vedrine, J. C., eds. (1994). "Catalyst Characterization: Physical Techniques for Solid Materials," Plenum Press, New York.
- Jansen, J. C., et al. (1994). "Advanced Zeolite Science and Applications (Studies in Surface Science and Catalysis, Vol. 85)," Elsevier, Amsterdam; New York.
- Meyers, R. A., ed. (1996). "Handbook of Petroleum Refining Processes," 2nd Ed., McGraw-Hill, New York.
- Moser, W. R., ed. (1996). "Advanced Catalysts and Nanostructured Materials: Modern Scientific Methods," Academic Press, San Diego.
- Pines, H. (1981). "The Chemistry of Catalytic Hydrocarbon Conversions," Academic Press, New York.
- Rase, H. F. (1999). "Handbook of Commercial Catalysts," CRC Press, Boca Raton.
- Sadeghlei (1995). "Fluid Catalytic Cracking Handbook," Gulf Publishing Company, Houston.
- Spitz, P. H. (1988). "Petrochemicals: The Rise of an Industry," Wiley Interscience, New York.
- Starks, C. M., Liotta, C. L., and Halpern, M. (1994). "Phase-Transfer Catalysis: Fundamentals, Applications, and Industrial Prospectives," Chapman and Hall, New York.
- Thomas, J. M. *et al.* (1997). "Principles and Practice of Heterogeneous Catalysis," VCH, New York.
- Van Santen, R. A., and Niemanstverdiert, J. W. (1995). "Chemical Kinetics and Catalysis (Fundamental and Applied Catalysis)," Plenum Press, New York.



Catalyst Characterization

Robert J. Farrauto

Melvin C. Hobson

Engelhard Corporation

- I. Physical Forms of Heterogeneous Catalysts
- II. Physical Properties of Catalysts
- III. Chemical Properties
- IV. Complementary Techniques

GLOSSARY

Active sites Microscopic locations of a catalyst where chemisorption and reaction to products occur.

Catalyst Substance that alters the rate of a chemical reaction without itself being consumed or generated.

Chemisorption Adsorption of reactants onto catalytic sites with comparable energetics and kinetics as a chemical reaction.

Dispersion Measured number of catalytic sites available to a probe gas compared with the total ideally present.

Pore size Approximate diameter of the microchannels within a porous material.

Pore volume Amount of void volume within a material.

Surface area Internal surface of a material accessible to physically adsorbed gas.

Washcoat High surface area oxide impregnated with catalytic species and bound to the walls in the channel of a monolithic structure.

THE CHARACTERIZATION of a heterogeneous catalyst is the quantitative measure of those physical and chemical properties of the catalyst assumed to be respon-

sible for its performance in a given reaction. Measurements include composition, active sites, and the external and internal structure of the solid material through which reactants and products must be transported during catalysis. Within the catalyst matrix itself (when there is an internal structure) five fundamental processes must take place (in order of occurrence): (1) diffusion or transport of reactant(s) to active sites through the pore structure of the catalyst, (2) chemisorption of reactant(s) onto active sites, (3) chemical reaction of chemisorbed species to produce product(s), (4) desorption of product(s) from active sites, and (5) diffusion or transport of product(s) through the pore structure. [Figure 1](#) is an idealized picture of these five fundamental steps. These steps are common to all catalysts with internal pore structures, whereas nonporous materials perform catalysis only on their external surface. These phenomena are referred to frequently throughout this article since the catalyst properties to be characterized determine the efficiency of the catalytic reaction. This review is divided into four sections. Section I describes many physical forms of heterogeneous catalysts used commercially. Sections II and III, physical and chemical properties, respectively, describe those techniques most important in characterizing

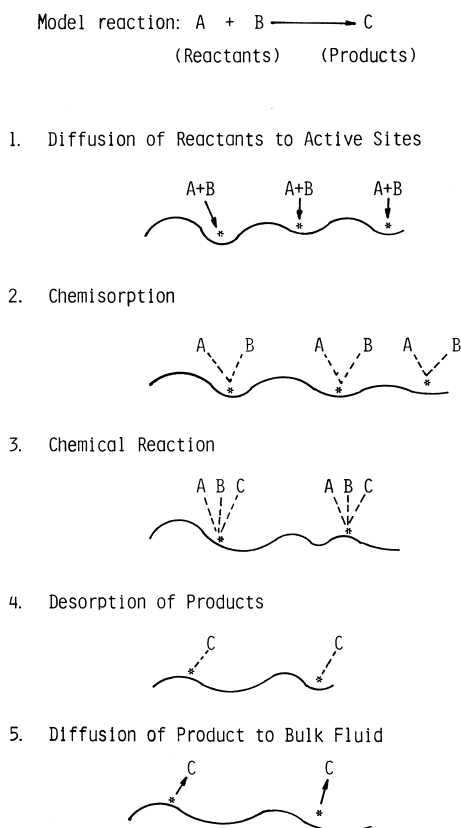


FIGURE 1 Schematic of the possible rate-controlling steps in a heterogeneously catalyzed reaction.

real industrial catalysts. The final section, Section IV, discusses nonroutine complementary techniques which when used in concert with those of Sections II and III provide fundamental property data. These, however, are not commonly used in practice in industry.

I. PHYSICAL FORMS OF HETEROGENEOUS CATALYSTS

A. Supported Catalysts

1. Powders

Powdered catalysts are used almost exclusively in slurry-phase reactions, in which the catalyst powder is mixed with a reactant. Vigorous agitation improves the contacting of reactant(s) and catalyst, and the rate of conversion of reactant to product(s) is monitored by suitable analytical techniques. The agitation is provided by an internal impeller; however, “rocking” or “shaking” reactors are also used. Separation of the catalyst from the reaction mix is usually accomplished by filtration, although sedimentation can also be used depending on the settling rate of the

catalyst. Reactions are typically carried out in batch autoclaves under pressure. Processes involving hydrogenation, alkylation, isomerization, and so on are commonly carried out in slurry-phase reactors.

A wide variety of catalytic materials are used as slurry-phase catalysts, most being metals supported on high surface area alumina, carbon, and silica (Fig. 2, label 3). Physical properties such as density are important since these catalysts must be suspended in the reaction mix. Since rapid agitation could lead to abrasion and attrition of the catalyst particles, strength is important.

Reactants and catalyst must be contacted; thus, a high external surface area of the catalyst (smaller particle size) is desirable to maximize reaction rate. The particle size of the catalyst must be optimized to permit filterability for ease of recycle while maintaining the high external surface area needed for maximum reactant–catalyst contacting.

The internal structure, comprising pores and surface area, is important for making the active catalytic sites accessible to the reactant molecules. The location of the active species is important for minimizing diffusional resistance since reactants must diffuse within the particle to the active sites and products must diffuse away. Finally, high catalytic surface area and high dispersion of active species are advantageous for maximum reaction rate and utilization of the catalytic components.

2. Particulates

Particulate catalysts are commonly used in fixed-bed reactors, in which the feed is passed through an immobilized bed of catalyst. Oxidation, hydrogenation–dehydrogenation, isomerization, alkylation, and hydro-treating are carried out in such reactors. Supported catalysts are composed of an active catalytic species dispersed throughout the support matrix. The supports take on many different sizes and shapes, all of which are determined by the reactor engineering. Spheres, extrudates, and tablets (Fig. 2, labels 2, 4, 6, and 7) are the most popular shapes; however, rings, stars, doughnuts, and others, are also used for specialized applications.

They are usually charged to a reactor of a fixed volume, and thus the bulk density influences the weight of catalyst present. Since these fixed-bed reactors can be very large (i.e., 15,000–20,000 lb per charge), a crush strength resistance of some minimum value is often specified. Large volumes of feed, frequently at high linear velocities, pass through the bed; thus, the resistance to abrasion must be considered.

All of the internal properties such as pore size, surface area, catalytic species location, and catalytic surface area are important since the five fundamental steps mentioned in the opening paragraph are operative.

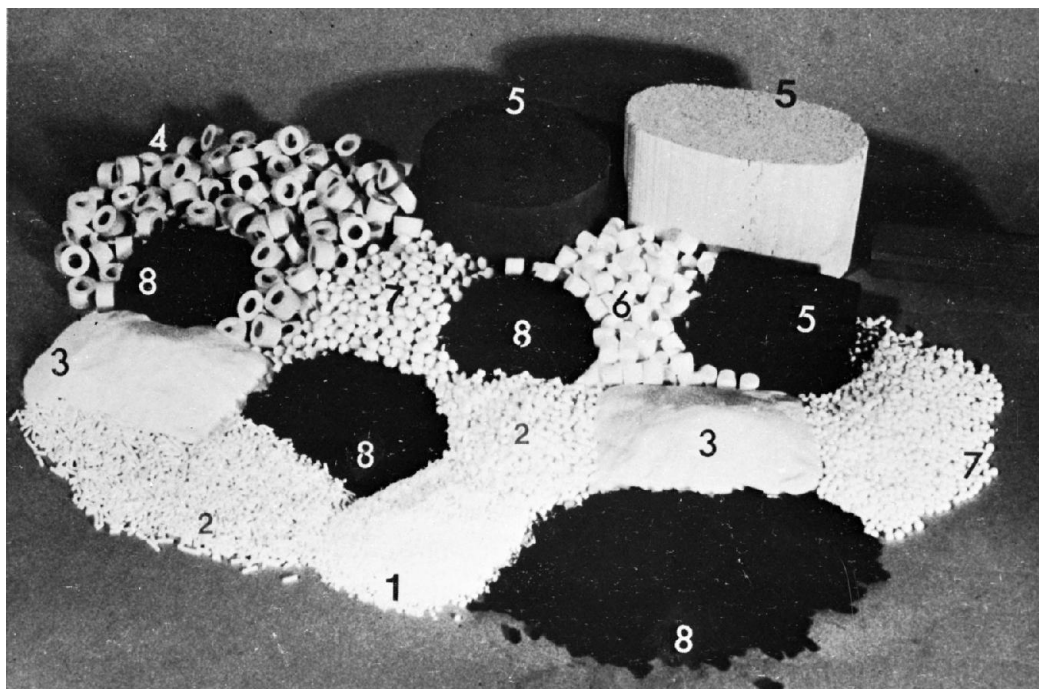


FIGURE 2 Some physical forms of heterogeneous catalysts. 1, Particulates; 2, extrudates; 3, powders; 4, rings; 5, monoliths; 6, tablets; 7, spheres; 8, carbon powders and particulates.

B. Unsupported Catalysts

1. Powders

Powders possessing relatively high surface area and active sites can be intrinsically catalytically active themselves. Powders of nickel, platinum, palladium, and copper chromites find broad use in various hydrogenation reactions, whereas zeolites and metal oxide powders are used primarily for cracking and isomerization. All of the properties important for supported powdered catalysts such as particle size, resistance to attrition, pore size, and surface area are likewise important for unsupported catalysts. Since no additional catalytic species are added, it is difficult to control active site location; however, intuitively it is advantageous to maximize the area of active sites within the matrix. This parameter can be influenced by preparative procedures.

2. Gauzes

Metal wires and screens are used as fixed-bed catalysts in which reactants are passed through the openings in the gauze, the size of which is defined by the mesh and wire diameter (see Fig. 10A). Gauzes composed of an alloy of platinum and rhodium catalyze the air oxidation of ammonia to nitric oxide, which is subsequently converted to nitric acid, and the production of hydrogen cyanide from ammonia, air, and methane. Formaldehyde production by

the oxidation of methanol is sometimes carried out with screens or gauzes of silver.

These catalysts are manufactured as smooth wires with no internal pores and then woven into gauze pads. Mechanical rigidity is important since the reactors are usually large in diameter (i.e., 4–12 ft) and are used in the reactor with minimum physical support. Furthermore, the conditions of operation are quite severe with respect to temperature and corrosion, and thus metallurgical integrity must be maximized. The most important properties are the purity of composition, wire diameter, and mesh size as well as mechanical strength.

3. Unsupported Particulate Catalysts

Unsupported particulates, like their powder counterparts, contain active sites without the addition of other catalytic species. Synthetic zeolites and $\text{SiO}_2\text{--Al}_2\text{O}_3$ catalysts used for cracking heavy oils to gasolines are catalytic due to their acid sites. They are produced by chemical reactions between the various components but can be found in nature. These materials are often modified by chemical techniques such as ion exchange; however, the impregnation techniques typical of dispersed catalysts are not used. Promoters can be added to enhance performance.

These materials are usually used in moving- or fluidized-bed reactors and thus are prone to severe attrition. Furthermore, because they are fluidized their

suspension properties, related to particle size and density, are important. Pore size is critical for shape-selective applications such as the dewaxing of lubricating oils or synthesis of molecules of a particular size.

Massive metals themselves are used as unsupported fixed-bed catalysts; for example, Raney nickel is used in a variety of hydrogenation reactions. The synthesis of ammonia from N_2 and H_2 is carried out with reduced massive iron containing minor amounts of promoters.

C. Monolithic Catalysts

1. Catalyzed Washcoats on Monoliths

A slurry of a high surface area oxide (Al_2O_3 , TiO_2 , SiO_2 , etc.) is deposited as a thin layer onto the channels of a ceramic or metal honeycomb. This washcoat is then made active by impregnation with catalytic species. The honeycombs vary in composition, cell density and shape, and wall thickness (Fig. 2, label 5). They must have sufficient surface porosity or roughness to allow the washcoat to adhere tightly. The overall geometry is dictated by the dynamics of the reaction of interest, but the most common use for honeycomb catalysts is for high-throughput gas reactions where pressure drop must be minimized such as in pollution abatement from both moving (auto exhaust) and stationary (chemical plants) sources. Other applications in the chemical industry are being pursued.

The catalyzed washcoat possesses internal structure similar to those of catalyzed powders and particulates, and hence the properties applicable to them are also important to the washcoat. In addition, washcoat adhesion plays a critical role since gas throughputs are extremely high and exfoliation can be a common problem.

The most widespread use of monoliths is for catalytic conversion of pollutants generated from the internal combustion engines of automobiles. Thus, the material must be mechanically strong to resist vibration and rapid temperature excursions.

II. PHYSICAL PROPERTIES OF CATALYSTS

A. Surface Area, Pore Size, and Pore Volume

Surface area, pore size, and pore volume are among the most fundamentally important properties in catalysis because the active sites are present or dispersed throughout the internal surface through which reactants and products are transported. The pores are usually formed by drying or calcining precipitates of hydrous oxides; however, some materials possess porosity naturally, as in the case of carbons, natural zeolites, and others. Raney nickel catalysts

are made porous by the selective leaching of an alloy constituent, usually aluminum. Combustible substances are incorporated into ceramics, which, when burned out, create pores in the host ceramic. Finally, during catalysis a material may become more porous by the volatilization or recrystallization of certain components, the most common example being PtRh (or PtPdRh) alloys used for the oxidation of ammonia to nitric acid, which becomes porous by the volatilization of platinum oxides during the reaction.

The size and number of pores determine the internal surface area. It is usually advantageous to have high surface area (high density of small pore sizes) to maximize the dispersion of catalytic components; however, molecules such as those present in heavy petroleum or coal-derived feedstocks may be so large that they are excluded from small pores. The pore structure and surface area must be optimized to provide maximum utilization of active catalytic sites for a given feedstock.

1. Gas Adsorption: Surface Area

The most common procedure for determining the internal surface area of a porous material, with surface areas greater than 1 or 2 m^2/g and up to $\sim 1200 m^2/g$, is based on the adsorption and condensation of N_2 at liquid N_2 temperature. The partial pressure of N_2 above the sample is gradually increased, and N_2 molecules are physically adsorbed onto the surface, approaching monolayer coverage (first steep portion of isotherm shown in Fig. 3A). Each adsorbed molecule occupies an area of the surface comparable to its cross-sectional area ($\sim 16.2 \text{ \AA}^2$). By measuring the number of N_2 molecules adsorbed at monolayer coverage, one can calculate the internal surface area. In practice, coverage beyond a monolayer occurs, and at high relative N_2 partial pressures, condensation of liquid N_2 in the pores occurs. The Brunauer, Emmett, and Teller (BET) equation describes the relationship between volume adsorbed at a given partial pressure and the volume adsorbed at monolayer coverage:

$$\frac{P}{V(P_0 - P)} = \frac{1}{V_m C} + \frac{(C - 1)P}{V_m C P_0}.$$

Here, P is the partial pressure of N_2 , P_0 the saturation pressure at the experimental temperature, V the volume adsorbed at P , V_m the volume adsorbed at monolayer coverage, and C a constant.

This equation can be linearized by plotting $P/V(P_0 - P)$ against P/P_0 , in which the slope is $(C - 1)/V_m C$, whereas the intercept is equal to $1/V_m C$ (Fig. 3B). The sum of the slope and intercept yields the reciprocal of V_m . The most reliable results are obtained at relative pressures (P/P_0) of between 0.05 and 0.3.

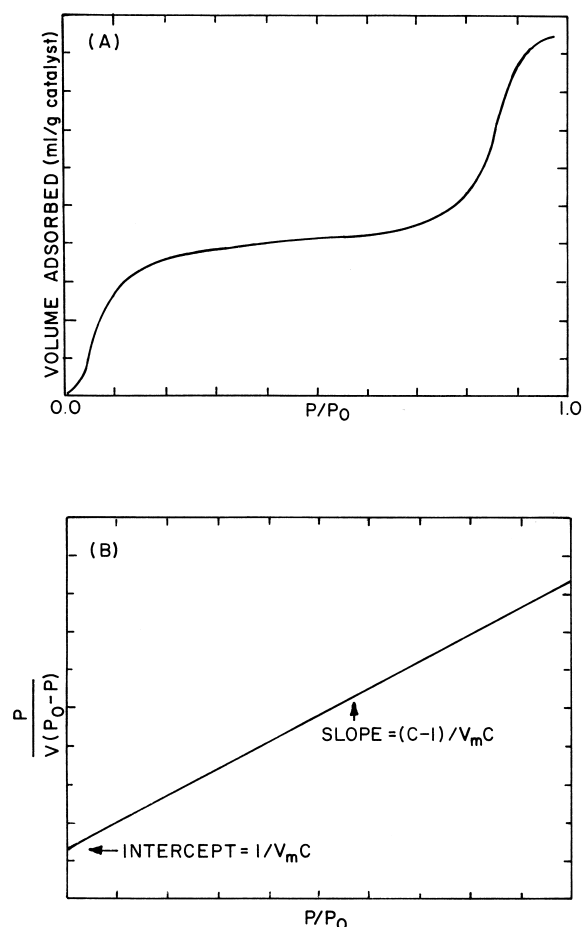


FIGURE 3 Measurement of surface area by the BET gas adsorption method. (A) Typical adsorption isotherm with a relatively flat curve in the region of monolayer adsorption. (B) Plot of the linear form of the BET equation between $p/p_0 = 0.05$ and 0.3 used to calculate the monolayer coverage V_m .

2. Gas Adsorption: Pore Size

The same equipment as that for measuring surface area can be used to determine the pore size distribution of porous materials with diameters of 20 to 500 Å, except that high relative pressures are used to condense N_2 in the catalyst pores. The procedure involves measuring the volume adsorbed in either the ascending or the descending branch of the BET plot at relative pressures close to 1. Capillary condensation occurs in the pores in accordance with the Kelvin equation,

$$\ln P/P_0 = -\frac{2\sigma V \cos \theta}{rRT},$$

where σ is the surface tension of liquid nitrogen, θ the contact angle, V the molal volume of liquid nitrogen, r the radius of the pore, R the gas constant, T the absolute temperature, P the measured pressure, and P_0 the saturation pressure.

Hysteresis in the adsorption–desorption isotherms (Fig. 4) is a common observation for supports with a large fraction of small pores. It results from desorption from the meniscus at the end of a filled pore. The vapor pressure above the liquid at the pore mouth defines the pore radius in the Kelvin equation. Therefore, it is the desorption branch of the isotherm that is preferred in calculations of pore size distributions.

3. Mercury Intrusion

The penetration of mercury into the pores of a material is a function of applied pressure. At low pressures mercury penetrates the large pores, whereas at higher pressures the smaller pores are progressively filled. Due to the nonwetting nature of mercury on oxide supports, penetration is met with resistance. The Washburn equation relates the pore diameter d with the applied pressure P :

$$d = \frac{-4\gamma \cos \theta}{P}$$

The wetting or contact angle θ between mercury and solid is usually 130° , and the surface tension of the mercury, γ , is 0.48 N/m. Pressure is expressed in atmospheres and d in nanometers (10 Å). This technique is satisfactory for pores down to 50 Å diameter; however, this is a function of the instrument capability. Maximum diameters measured are usually 10^6 Å .

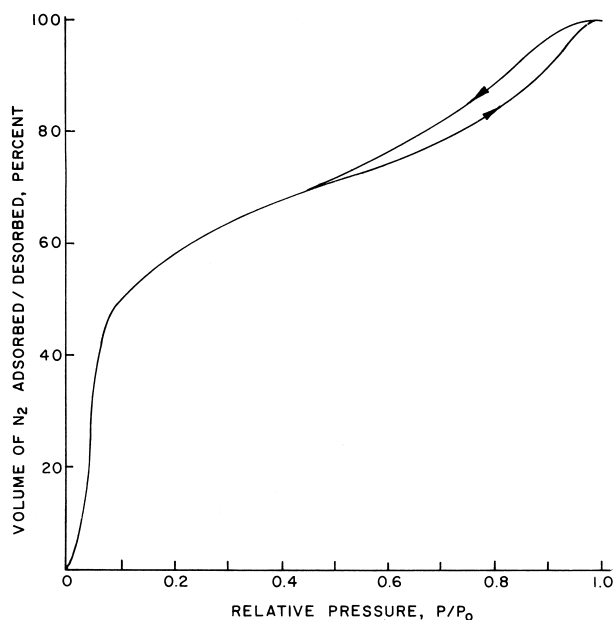


FIGURE 4 Nitrogen adsorption and desorption isotherms at 78 K. Pore size distributions in the micropore range are calculated from the isotherms using the Kelvin equation.

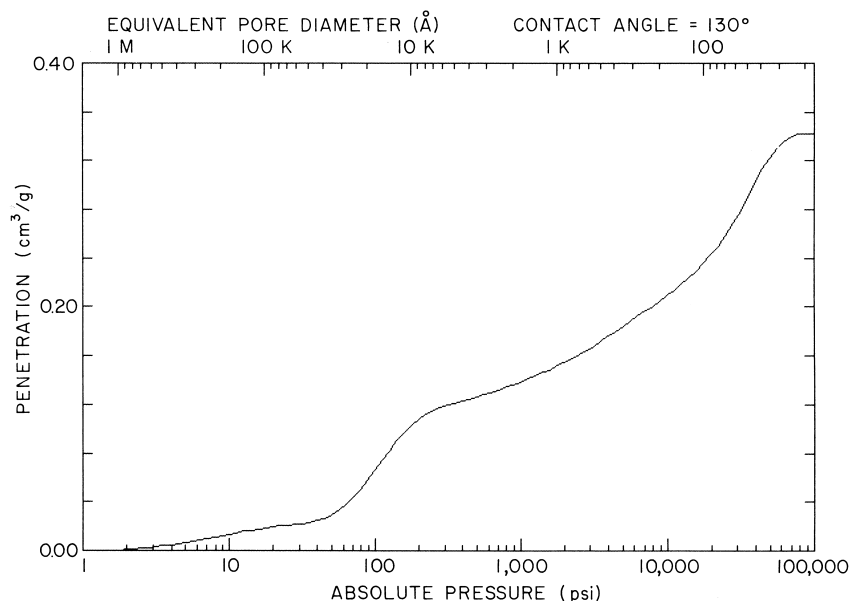


FIGURE 5 Pore size distribution by the mercury intrusion method.

Typical pore size distribution data are shown in Fig. 5, where the integral penetration of mercury into the pores is plotted as a function of applied pressure. The calculated pore diameters in angstroms are shown across the top. The integral curve clearly shows a bimodal pore distribution with mean pore diameters at 20,000 and 50 Å. The latter is at the lower limit of the technique. A nitrogen desorption isotherm is required to obtain an accurate measure in the region below 100 Å.

B. Particle Size Distribution

1. Powders

Powders vary dramatically in particle size on the basis of their origin. It is common for catalyst manufacturers to classify powders in order to assure users of consistency from batch to batch since suspension, settling rates, filtration, and performance in slurry-phase reactions are all dependent on particle size. The effect on suspension, settling rates, and filtration is obvious. However, factors that favor these are unfavorable for kinetics. For reactions controlled by transport rates from the bulk fluid to the surface of the catalyst, the overall reaction rate is a strong function of geometric surface area and thus is favored by small particles. Pore diffusion resistance is also minimized by smaller particles since reaction paths to active sites are smaller. The only mode of reaction control not influenced by particle size is for those reactions in which rate is controlled by reaction at active sites. Therefore, a compromise for optimum filtration and maximum reaction rates must be made.

Sieves of various mesh sizes have been standardized, and thus one can determine particle size ranges by noting the percentage of material, usually based on weight, that passes through one mesh size but is retained on the next finer screen. Sieves are stacked with the coarsest on top and the underlying screens progressively finer. A precise weight of catalyst material is added to the top screen. The stack of sieve is vibrated, allowing the finer particles to pass through coarser screens until retained by those screens finer in opening than the particle size of the material of interest. Each fraction is then weighed and a distribution determined.

This method is reliable only for particles larger than $\sim 40 \mu\text{m}$. Below this, sieving is slow and charging effects influence measured values. Sophisticated instrumentation is available for measuring the distribution of finer particles. Methods include light scattering, image analysis, sedimentation, centrifugation, and volume exclusion. An example of volume exclusion is shown in Fig. 6, where data obtained with a Coulter counter are presented. In this method the powder is suspended in an electrolyte and pumped through a tube containing a small orifice. An electric current passes through this tube, and as individual particles pass through the orifice a fraction of the current is interrupted. This fractional change in current is a measure of particle size. The magnitude of the change in current flow is subdivided over the range of sizes limited by the size of the orifice. The particle diameters are calculated on the equivalent sphere of the excluded volume, and, assuming constant density for the particles, the results are commonly recorded as weight percentage as a function of

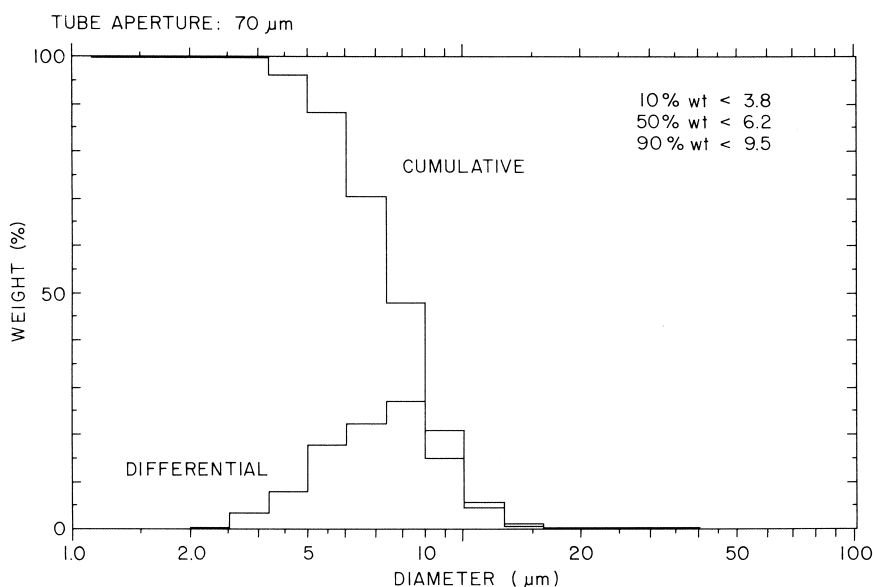


FIGURE 6 Particle size distribution of a γ -alumina powder measured by an excluded volume method, in this case involving a Coulter counter.

incremental particle size (Fig. 6). Obviously, irregularly shaped particles will introduce deviations into the results based on a spherical model. The wood carbon shown in Fig. 8 is an example of an irregularly shaped catalyst support for which particle size distribution measurements are important but difficult to measure satisfactorily.

Other techniques, such as light scattering and sedimentation, are also sensitive to particle shape. The direct observation of particles by a scanning electron microscope and distribution measurements by image analysis would appear to overcome many of the problems associated with the various other techniques, but problems of describing irregularly shaped particles remain. A universal particle size descriptor has not yet been developed. The technique adopted and the results obtained are most useful when empirical correlations with the end use can be made.

2. Particulates

The particle size distribution influences the packing of particulate catalysts in fixed-bed reactors and thus affects such process parameters as flow rates, reactant–catalyst contacting, temperature control, and pressure drop, all of which influence product distribution and yields. Large catalyst particle sizes favor low pressure drop but may not permit proper contacting of feed with catalyst, resulting in bypass. Reactions controlled by bulk mass transfer of reactants to the external surface are favored by smaller particle sizes to maximize geometric area. Rates of pore-diffusion-controlled reactions are also enhanced by decreasing particle size. An optimum must be met, therefore, between reaction kinetics and reactor and process design.

Screens of precisely calibrated mesh sizes and openings are the principal devices for measuring the distribution of sizes for particulate catalysts. Procedures are discussed in Section II.B.1.

3. Washcoat Thickness

Washcoat thickness is analogous to particle size in that reactants must penetrate its pore structure and interact with the dispersed active sites. The products produced must diffuse through the structure and out into the bulk gas. This phenomenon differs from that involving a particle in that only the gas–solid washcoat surface is available since the other side is bonded to the wall of the monolith.

Optical microscopy is the method used most frequently to obtain thicknesses directly. A portion of monolith is mounted in epoxy and sliced to obtain a cross section. The contrast between washcoat and monolith is sufficient to permit thickness measurements to be made optically. A typical cross section of a washcoat on a ceramic auto exhaust monolith is shown in Fig. 7.

C. Mechanical Strength

1. Crush Strength of Single Pellet

Particulates packed in reactor beds are subjected to the static pressures of the bed height and thus must be sufficiently strong to resist crushing. Monoliths, particularly when used in a stacked mode, for example, in stationary pollution abatement, must resist crushing axially. For vehicular use, for example, auto exhaust and ozone

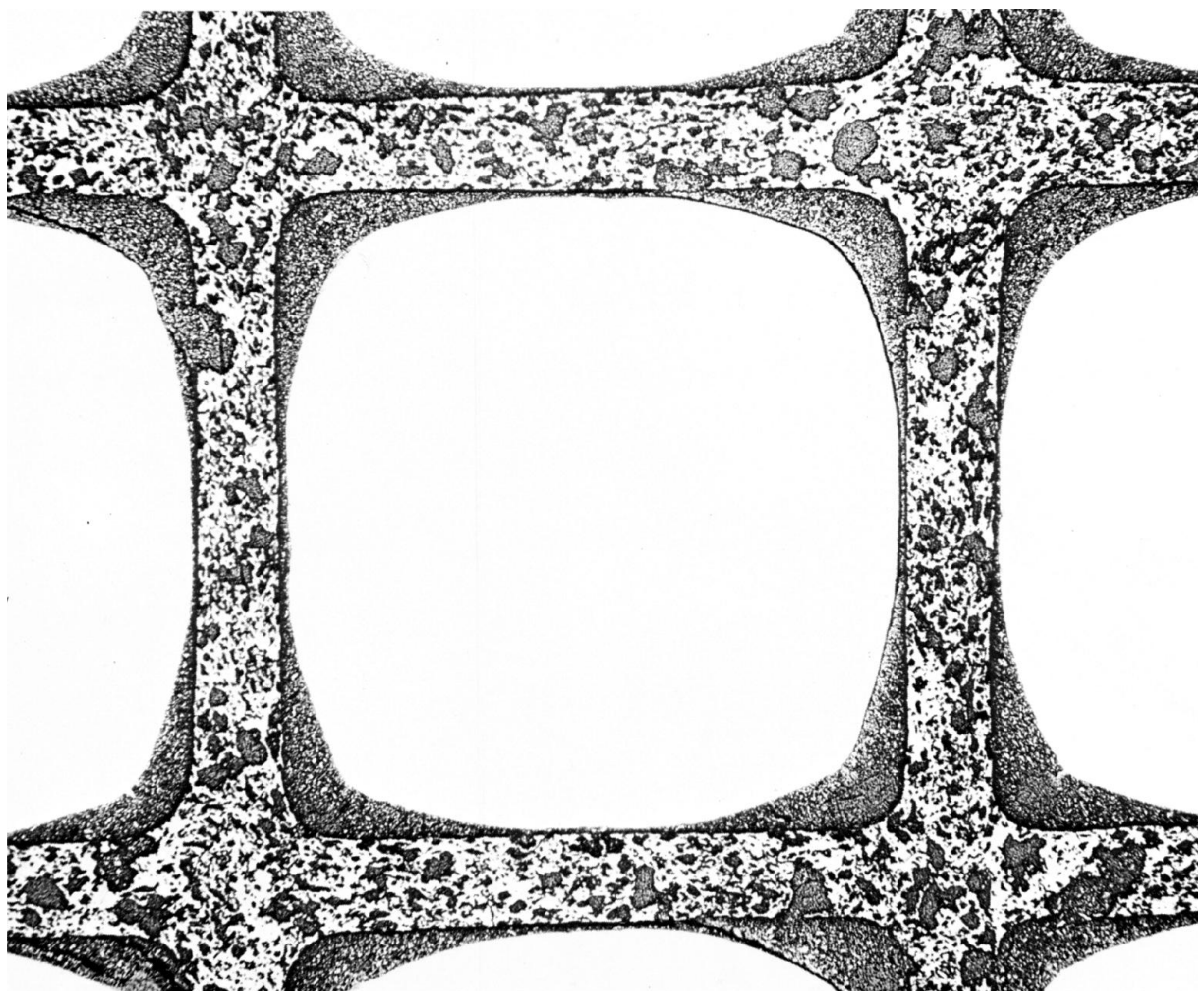


FIGURE 7 Optical photomicrograph of a cross section of a honeycomb catalyst. The thickness of the granular material, the washcoat, can be measured with a calibrated reticle in the microscope eyepiece.

abatement in aircraft, resistance to vibration as well as radial strength is important.

Measurement for tablets, spheres, extrudates, monoliths, and so on, is quite simple. A single particle or unit representative of the lot is placed between parallel plates of a device capable of exerting compressive stress, and the force necessary to crush the material is noted. Tablets, extrudates, or monoliths can be placed within the plates so as to measure axial or radial crush strength. Naturally, a sufficient number of particles must be tested to obtain proper statistics. The method cannot be reliably used for granular or other materials with irregular shapes.

2. Attrition and Abrasion

Catalyst particles in fluid-bed reactors such as those used in heavy oil cracking are subjected to potential abrasion

by collision with each other or the internal surfaces of the reactor. Thus, abrasion resistance is a key catalyst property.

Catalyst material (tablets, extrudates, spheres, or irregularly shaped catalysts) are charged to a drum, which rotates for a given set of time, and the “fines” produced are measured. The drum usually contains a single baffle, and its inner surface must be smooth. Powder attrition can be measured using particle size distribution techniques, as described in Section II.B.1. Monolithic catalysts are subjected to high gaseous flow rates, and hence erosion of the washcoat from the surface of the monolith is a likely cause for concern. The most common technique is simply to subject the washcoated monolith to a jet of gas simulating the linear velocities anticipated in use and note the weight change of the catalyst due to erosion. Naturally, one can also collect the attrited particles and correlate

their weight with the weight of washcoat initially bound to the monolith.

Monolithic materials are frequently subjected to substantial thermal gradients (thermal shocks) in startup and shutdown, as in auto exhaust emission control. Frequent thermal shocking causes the washcoat to lose adherence due to the expansion difference between it and the monolith. This is most pronounced when the monolith is metallic.

One can evaluate this phenomenon by periodically subjecting the catalyzed material to gaseous flows at temperatures anticipated in service and noting weight losses. Monoliths can be mounted on a rotating “carousel” that moves in and out of streams of heated gas.

D. Density

1. Bulk or Packing

Particulate catalysts are usually sold by weight but are charged to a reactor by volume. Thus, the density of the support has a strong impact on the economics of the process. Fluidization in moving-bed reactors is also dependent on catalyst density. The density of powders affects the extent to which they can be suspended and eventually settled in slurry-phase reactors.

Samples of powders or of formed particles such as pellets, spheres, or extrudates are first dried at a temperature sufficient to remove moisture or organic contaminants ($\sim 400^\circ\text{C}$). The material is cooled and vibrated while being poured into a cylinder of known volume and weight. The volume occupied by the catalyst is noted along with its weight. The weight of the catalyst divided by its volume is its apparent packing density. In place of a vibrator one could also use a tapping device and thus obtain the tapped apparent packing density.

2. Skeletal

The skeletal density is representative of the solid material itself, excluding its porosity. The bulk volume of catalyst minus its pore volume and the interparticle volume between discrete particles (V_I) is the true skeletal volume. One calculates this term by

$$d_{\text{skeletal}} = M/V_{\text{skeletal}}$$

where M is mass and $V_{\text{skeletal}} = V_{\text{bulk}} - V_{\text{pore}} - V_I$. The pore volume is determined by the mercury intrusion method (Section II.A.2); however, only pores greater than $\sim 30 \text{ \AA}$ are measured (for an instrument with a 60,000-psi capability). The pore volume that includes pores less than $\sim 30 \text{ \AA}$ must be determined by cumbersome gas displacement techniques.

E. Morphology

1. Surface Texture via Electron Microscopy

Electron microscopic examination of catalyst materials, particularly those containing natural components, permits the identification of their origin. For example, carbons utilized as supports for precious metals in a wide variety of slurry-phase and fixed-bed reactions can be derived from a large number of naturally occurring sources (Fig. 8). The shape, morphology, and composition are useful properties for determining their origin.

Edges or surface irregularities on particulate catalysts used in fluid- or fixed-bed applications are susceptible to attrition and erosion during reaction. The morphology of a typical fluid-bed cracking catalyst is shown in Fig. 9. The surface of spheres, extrudates, and tablets is relatively free of topological features, and thus physical losses are usually not a serious problem.

The texture of PtRh gauzes before and after use in nitric acid production shows significant morphological changes due to vaporization and deposition of PtO_x species (Fig. 10C). The surface, as examined by electron microscopy, shows a swelling or sprouting effect that increases the surface area of the gauze during the early stages of use.

2. Adhesion of Washcoats to Monoliths by Optical Microscopy

Refractory high surface area oxides are deposited from slurries onto the walls of the channels that make up monoliths in order to provide an adequate surface area to support the active catalytic species. Washcoats such as Al_2O_3 and TiO_2 are commonly used for pollution abatement applications (auto exhaust, stationary NO_x abatement, etc.) where the monolith is usually a ceramic. Metal monoliths are finding increasing use; however, they represent only a small percentage of the total monoliths used. Optical microscopy enables one to see that the catalyzed washcoat follows the contour of the ceramic surface. Figure 7 shows the Al_2O_3 washcoat–ceramic interface for a typical auto exhaust catalyst. In this case, no evidence of loss of adhesion between washcoat and ceramic can be seen.

F. Location of Catalytic Species Within a Support

1. Electron Microscopy

For reactions that are pore diffusion controlled, it is advisable to locate the active catalytic species close to the fluid–solid interface in order to decrease the diffusion path of reactants and products. This applies to all forms

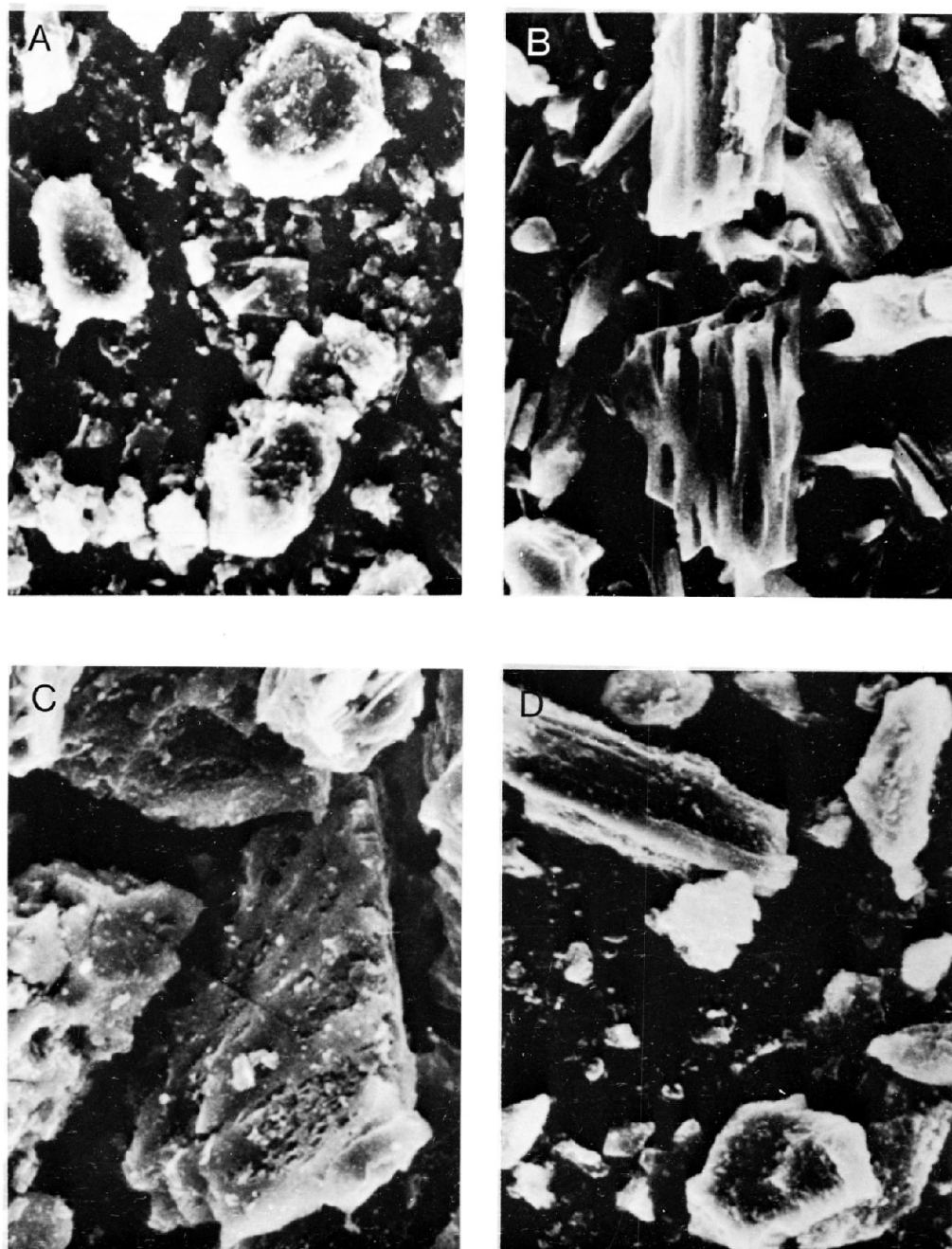


FIGURE 8 Typical morphology of carbon particles derived from various natural sources. A, Coal; B, wood; C, coconut; D, peat.

of catalysts, from powders to catalyzed washcoats. Uniform distribution of catalytic species is advisable for reactions controlled by chemical kinetics free of diffusional limitations.

If a process feed contains known catalyst poisons, such as lead compounds in the exhaust of automobile emissions, it may be wise to locate the active species more

deeply into the washcoat to avoid undesirable reaction between the poison and active catalytic component (i.e., platinum, palladium, or rhodium). In this way the poisons are deposited on the outside periphery of the catalyst before they reach the active sites. A similar situation exists for those cases where abrasion or surface erosion is expected. Obviously, if all the catalyst is located at the

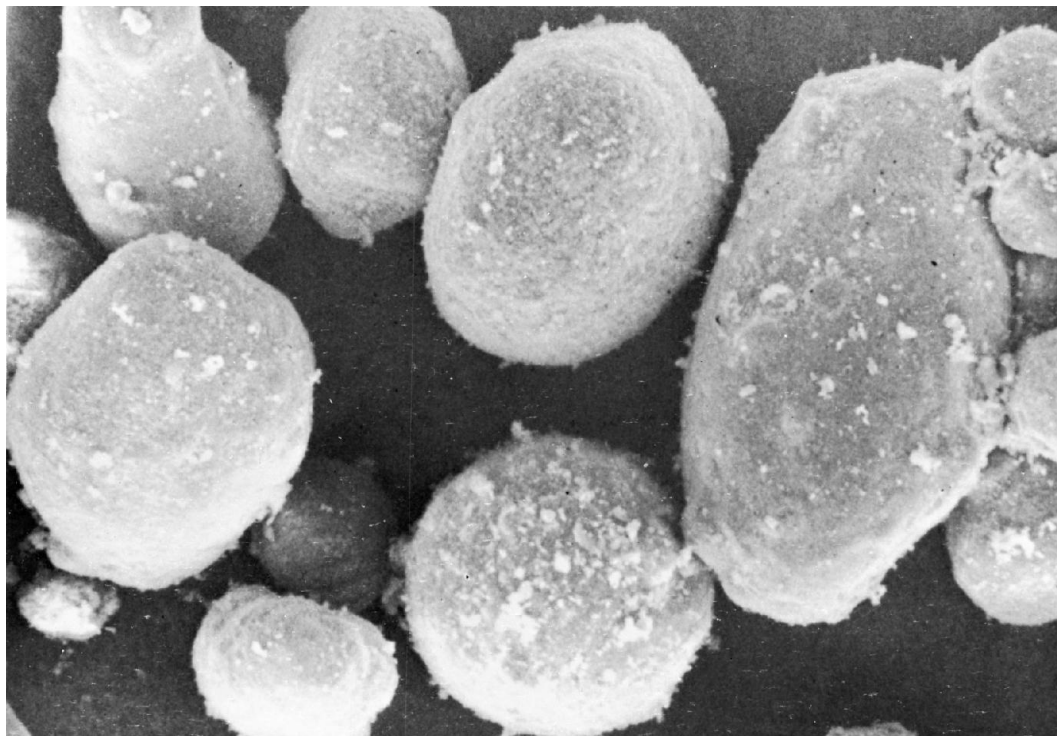


FIGURE 9 Typical morphology of a fluid cracking catalyst, 325 $\times$ magnification.

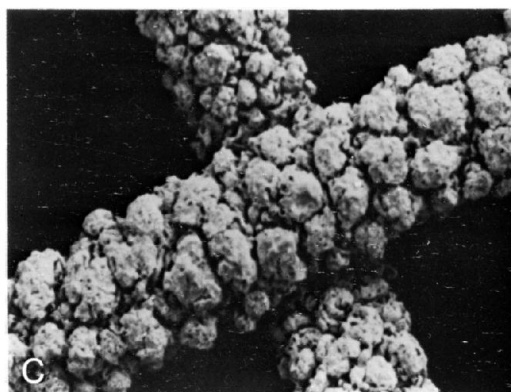
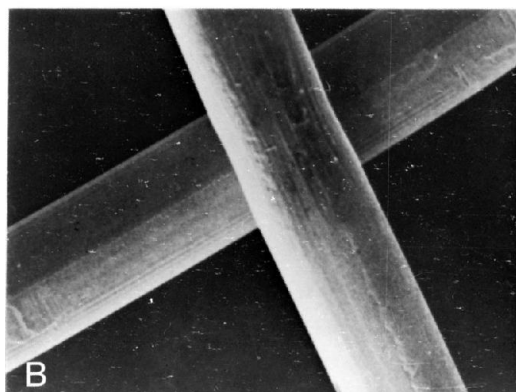
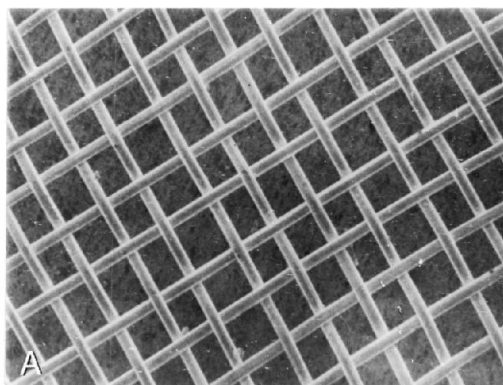


FIGURE 10 Platinum-rhodium wire gauze for ammonia oxidation. A, 20 $\times$ magnification of a fresh gauze; B, 300 $\times$ magnification of a fresh gauze; C, 300 $\times$ magnification of a used gauze.

fluid–solid interface, it is likely to be lost in an abrasive environment.

The optimum catalyst location must be determined by a delicate balance of the kinetics, which will control the overall reaction rate, and anticipated deposits or abrasive components contained in the feed.

Transmission and scanning electron microscopes and electron microprobes are electron optical instruments with the same basic features, including an electron gun with a variable high-voltage supply, magnetic lenses to focus the electron beam, and detectors to record the images or characteristic X-ray emissions from the sample. All these features are packaged in a high-vacuum system capable of reaching at least 10^{-5} torr. Transmission and scanning electron microscopes and the electron microprobe each exhibit capabilities that complement each other. The essential operating features of the STEM (an instrument combining scanning and transmission capabilities) will be described since it embodies the important characteristics of each of these instruments.

The basic electron optical system consists of an electron source, an array of magnetic lenses to collimate the electron beam, objective lenses, and a projector lens that focuses the images at the focal plane. The collimated electron beam passes through the specimen, is focused by the condenser lens on the back focal plane, and is magnified by intermediate lenses and finally the projector lens to form the final image. With the addition of deflector coils to the magnetic lens system, the electron beam can be rastered across a small area of the specimen. Defocusing of the objective lens increases the depth of field of the image in the transmission mode. Compared with that of optical microscopes the depth of field is much greater before any loss of resolution is observed. When the electrons are brought into focus in the back focal plane by the objective lens, an electron diffraction pattern is obtained. The diffraction pattern provides structural information of crystalline materials equivalent to X-ray diffraction patterns. The crystal structure and consequently identification of very small crystallites can be obtained by this technique, an important measurement in catalytic research.

In the scanning mode the electron beam focused on the sample is scanned by a set of deflection coils. Backscattered electrons or secondary electrons emitted from the sample are detected. As the electron beam passes over the surface of the sample, variations in composition and topology produce variations in the intensity of the secondary electrons. The raster of the electron beam is synchronized with that of a cathode ray tube, and the detected signal then produces an image on the tube.

In the transmission mode a thin sample, usually prepared by a microtome, is subjected to a beam of electrons, and those transmitted are noted. The dark spots on the pos-

itive of the detecting film correspond to dense areas in the sample, which inhibit the transmission of electrons. These dark spots form the outline of metal particles or crystallites, and hence their sizes can be determined. Electron diffraction can also be conducted, allowing the determination of the particle structure.

The electron microprobe is similar to the scanning electron microscope; however, its primary function is to detect characteristic X-rays produced by the electron beam interaction with the specimen. The X-ray emissions can be used to determine the elemental composition of the specimen quantitatively and the location of a particular element within the morphology or topological structure of the specimen.

2. Electron Microscopy: X-Ray Analysis

Both the electron microprobe and scanning electron microscope have proved to be very useful in determining the location of metal in catalyst particles. The morphology and metal location of a typical carbon-supported metal catalyst are shown in Fig. 11. The carbon particles are potted in epoxy resin and then cut, ground, and polished to provide a cross-sectional surface to be examined. From the morphology shown in Fig. 11, the particles are a wood-based carbon with pores typical of this type. The palladium X-ray maps (Fig. 11B) clearly outline the edge coating of palladium on the carbon particle. The coating penetrates no more than $15\text{ }\mu\text{m}$ into the interior of the carbon particle, which is $\sim 50\text{ }\mu\text{m}$ in diameter.

The location of the metal in a particle can also be determined by running a line scan to detect the X rays characteristic of palladium. As the electron beam moves across the particle, the detector signal is integrated and displayed on a cathode ray tube as intensity versus electron beam position. The resulting signal is shown in Fig. 11C.

Metal location is but one of a number of applications for scanning electron microscope studies in catalysis. Other applications are the study of the morphology of platinum–rhodium gauzes used in the oxidation of ammonia and the poisoning of catalysts, in which the scanning electron microscope results show the location of poisons such as compounds containing sulfur, phosphorus, heavy metals, or coke relative to the location of the catalytic components.

III. CHEMICAL PROPERTIES

A. Chemical Composition

1. Elemental Analysis

The precise nature of active sites is still the subject of considerable research; however, there is a huge body of data relating various metals, metal oxides, and compounds that,

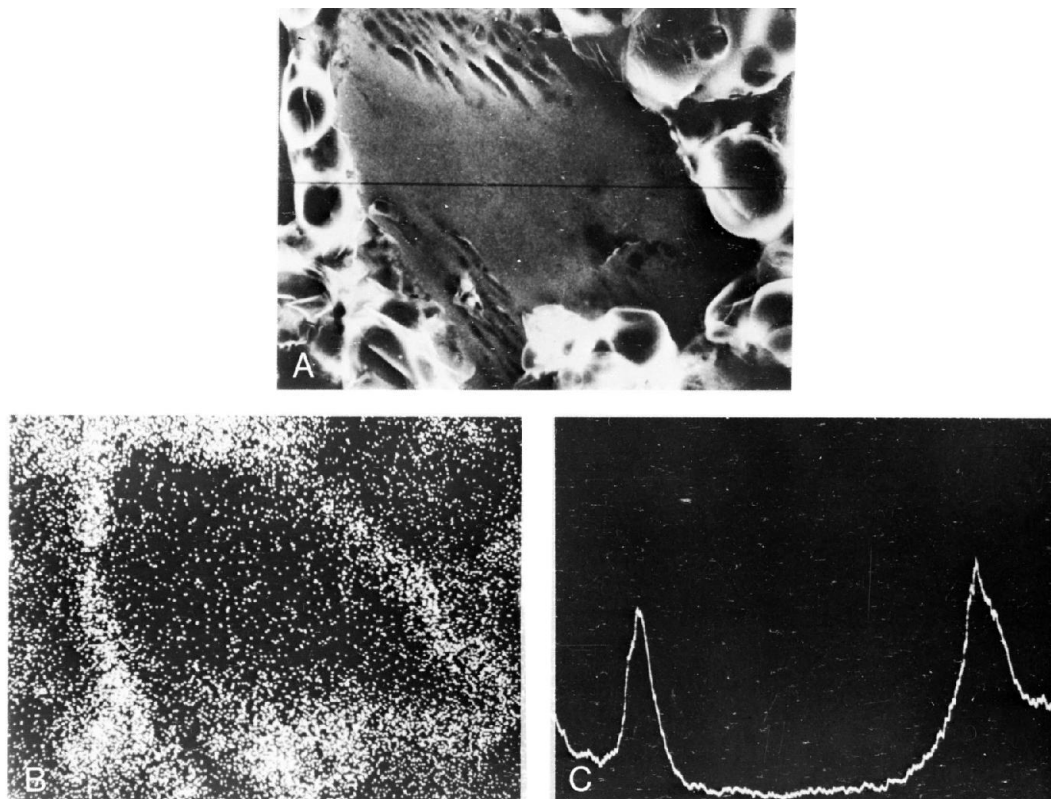


FIGURE 11 An edge-coated palladium-on-carbon catalyst. A, Photomicrograph of an irregularly shaped particle of wood carbon. B, Palladium map showing the palladium concentrated on or near the surface of the particle. C, Palladium line scan showing the distribution of the palladium at the edge of the carbon.

when present in or as a catalyst, generate active sites. The mere presence of these elements or compounds does not ensure catalytic activity since activation procedures must often be used. Heat treatment procedures can lead to the formation of a compound that generates active sites for a specific reaction, for example, copper chromites, bismuth molybdates, cobalt oxides, and PtO_2 . Activation can involve the reduction of an oxide to its metal, for example, in the reduction of NiO , Fe_3O_4 , and PdO . Acidic sites can be increased in catalysts in a variety of ways including ion exchange, addition of halides, impregnation with acids, or simply calcining. Although sulfur compounds are often catalyst poisons, they can be used to control selectivity or activity, for example, in hydrodesulfurization or in naphtha reforming.

The common denominator of all catalysts and activation procedures is the chemical composition necessary to generate active sites. The proper combination of chemical elements is essential in most catalysts for optimum performance. More often than not, small amounts of promoters ($\sim 0.1\%$ or less) or impurities can influence activity, selectivity, and life.

The ability to chemically analyze a catalyst properly is the subject of many specifications, which must be met by

catalyst suppliers. Commonly, catalysts are prepared using Al_2O_3 , SiO_2 , or carbon as supports. These materials are derived from raw materials, which contain various impurities that are usually detrimental to performance and thus must be removed. Impurities such as alkali and alkaline earth compounds, if used in excess, act as fluxes, causing sintering or loss of surface area in Al_2O_3 . The same impurities when added in the proper amount can enhance stability against sintering or in some cases improve selectivity. Hence, it is not just the presence of impurities, but the manner in which they have been introduced that is important. Obviously, one has greater control when starting with a relatively pure material to which can be added predetermined amounts of promoters.

The procedures used to analyze catalysts quantitatively are no different than those for any other chemical material. Often special procedures are needed to dissolve catalysts in preparation for analysis, particularly refractory materials such as certain noble metals and ceramics. The reader should refer to the large body of analysis information available for analyzing the chemical composition of a catalyst of interest.

Energy-dispersive analysis can be used to perform spot analysis within a catalyst. The electron beam can be

focused on a particle or area whose analysis is desired, and the X rays characteristic of the elements present are measured. This is a common method for poison analysis in a selected area of the catalyst. It should not be a substitute for a total analysis since only a small area is analyzed.

2. Analysis by X-Ray Diffraction

Provided that a material is sufficiently crystalline to diffract X rays and is present in an amount greater than $\sim 1\%$, X-ray diffraction (XRD) can be used for qualitative and quantitative analyses. The principle of this technique is that crystal structures possess planes made by repetitive arrangements of atoms, which are capable of diffracting X rays. The angles of diffraction differ for the various planes within the crystal, and thus every compound or element has its own somewhat unique diffraction pattern. The differences in these patterns, therefore, allow the differentiation of various structures within the catalyst. Figure 12 shows the XRD pattern of a methanol synthesis catalyst composed of CuO, ZnO, and Al₂O₃.

The compounds making up the catalyst sample can be clearly identified in the XRD pattern. Cupric oxide produces the peaks labeled C, zinc oxide the peaks labeled Z, and γ -alumina the peaks labeled A in Fig. 12. Not only does the XRD pattern qualitatively identify the phases present in the catalyst, but the quantity of each phase can be determined by measuring the area under selected diffraction peaks relative to a standard. An example of quantitative analysis by XRD is found in the ASTM Standard Procedure D3906-80 for NaY zeolite in a cracking catalyst.

Synthetic zeolites used for hydrocarbon cracking or isomerization reactions often have crystalline and amorphous components. This can be intentional, caused by the use of a binder, or could result from incomplete reaction during the production of the zeolite. Therefore, the percentage of

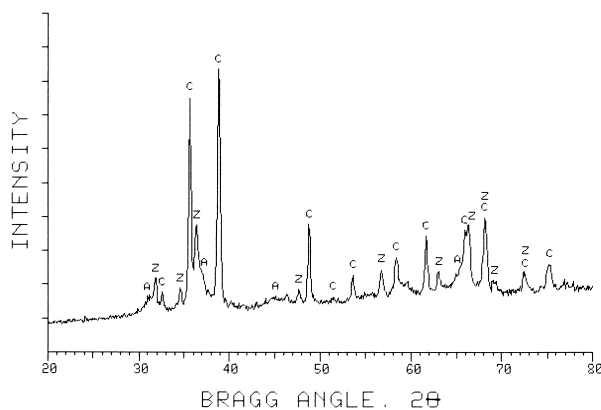


FIGURE 12 X-Ray diffraction pattern of a mixed oxide catalyst for methanol synthesis. The peaks marked C are cupric oxide; Z, zinc oxide; A, γ -alumina.

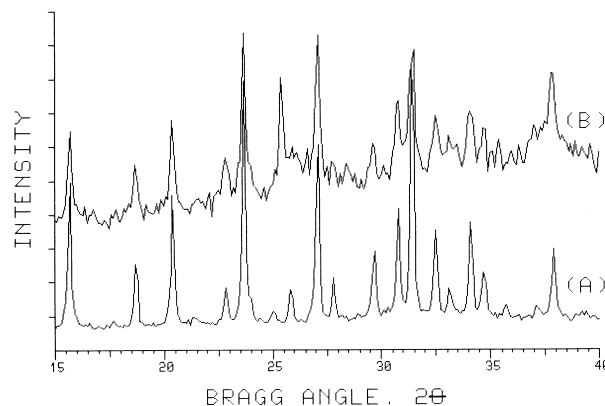


FIGURE 13 X-Ray diffraction patterns of a standard NaY zeolite (A) and a typical cracking catalyst containing the zeolite (B).

zeolite present is important for quality control as well as for defining the temperature limits of crystalline structure collapse. Figure 13 shows the XRD patterns of an NaY zeolite and a typical cracking catalyst. The sum of the intensities of eight peaks at 15.7, 18.7, 20.4, 23.7, 27.1, 30.8, 31.5, and 34.2° 2θ are used to compare intensities. The ratio of intensities of the zeolite peaks in the cracking catalyst pattern relative to the comparable peaks in the standard NaY yields the fractional amount of NaY present.

B. Structural Analysis

1. Crystallinity

The origin of the active site is based on the arrangement of metals, metal oxides, or multicomponent compounds. The previous section discussed the importance of the presence of elements and compounds, whereas this section indicates the importance of the structures made by these elements.

Figure 14 shows the XRD patterns of two Al₂O₃ structures, γ -Al₂O₃ and α -Al₂O₃. The former (Fig. 14B) is the high surface area, lower temperature structure, whereas the latter (Fig. 14A) is produced at high temperatures and has low surface area.

A major limitation in the use of XRD analysis for heterogeneous catalysts is that, below crystallite sizes of 30 to 50 Å, a well-defined X-ray pattern will not be obtained. Materials with crystallites smaller than that which is detectable are more precisely called amorphous since they possess no long-range order to diffract X-rays. Structures in this class, which are quite common for freshly prepared catalysts, must be characterized by other techniques. One such example is the increasing use of Si-29 and Al-27 NMR to provide structural information on these species in zeolites. Significant changes in the acidic properties of zeolite catalysts occur with changes in the Si/Al ratio and, thus, the cracking activity of FCC catalysts. NMR provides a means of determining this ratio in the presence

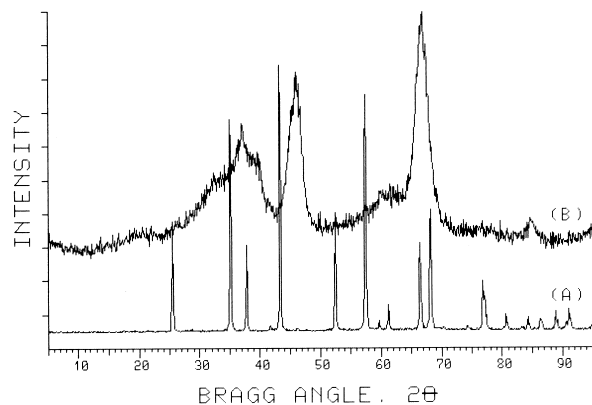


FIGURE 14 X-Ray diffraction patterns of crystalline and amorphous forms of aluminum oxide. Pattern A is the highly crystalline α - Al_2O_3 formed at high temperatures from B, the amorphous γ - Al_2O_3 phase.

of amorphous alumina and silica structures of the more common zeolites such as Y and mordenites.

2. Oxidation State

A very convenient method for approximating the oxidation state of a component within a compound is to reduce or oxidize a sample in a controlled environment and measure the weight change by microbalance techniques such as thermal gravimetric analysis. The catalyst weight (0.1–0.5 g) is equilibrated at a particular temperature in a stream of inert gas. A reactive gas, such as H_2 for a reduction or O_2 for an oxidation, is introduced to the inert gas stream and the associated weight changes monitored. The weight change associated with reduction or oxidation can give information regarding the valence state of the catalytic component. By comparison with bulk chemical

analysis, one can determine the extent of oxidized species present. Figure 15A shows a weight versus time profile for the reduction of magnetite, Fe_3O_4 , to the metal. A carefully controlled reduction of magnetite produces the iron catalyst used in the synthesis of ammonia from nitrogen and hydrogen. The apparent oxidation state can be estimated from the oxygen weight loss noted. Reoxidation (Fig. 15B) also enables one to approximate the oxidized state of iron by noting the weight gain due to oxygen reaction with the metal. The sample was partially reduced initially, as evidenced by the smaller weight loss indicated in curve A compared with the full oxidation shown for curve B. The temperature (dashed line) was programmed to increase linearly at $10^\circ\text{C}/\text{min}$.

An alternative technique based on the same principle is called temperature programmed reduction. The TPR apparatus is similar to a gas chromatograph. A stream of reactive gas passes through a small bed of sample and the composition of the gas is monitored by a thermal conductivity detector. The temperature of the catalyst is raised at a linear rate and the detector measures the consumption of the reactive component of the gas as a function of temperature. Reduction and oxidation of the catalyst, desorption of chemisorbed gases, and catalyzed reactions can be studied quantitatively by this technique.

C. Dispersion of Catalytic Species

1. Chemisorption

One of the most frustrating facts facing the catalytic scientist is that often when a structure has a definite XRD pattern and thus can be structurally well characterized, it usually has less than optimum activity. Large crystals possess many subcrystals, which diffract X rays and thus

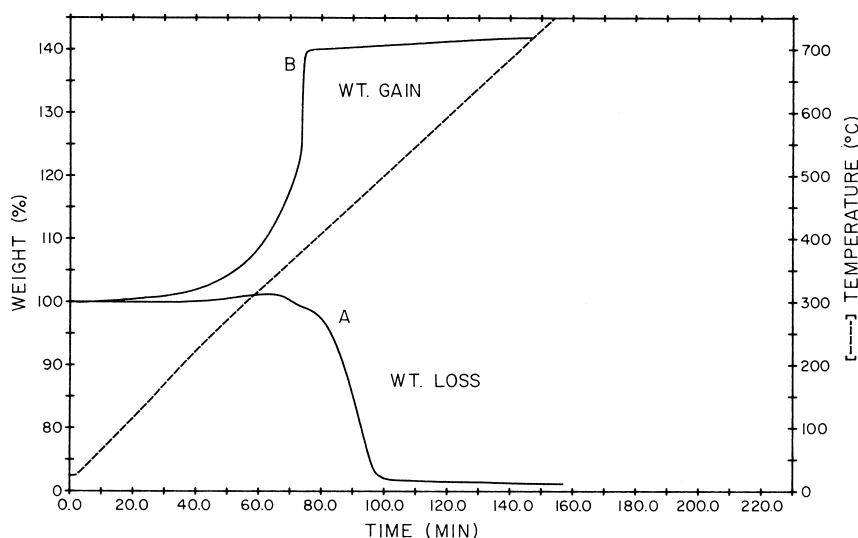


FIGURE 15 Thermogravimetric analysis of magnetite, Fe_3O_4 . A, Reduction in flowing H_2 at constant rate of temperature rise; B, reoxidation of the reduced sample.

generate easily read patterns; however, the subcrystals are buried within the crystal, making them inaccessible to reactant molecules. Frequently the purpose of the preparation technique is to disperse the catalytic components in such a way as to maximize their availability to reactants. When this is done effectively, the crystal structure is composed of fewer subcrystals, and thus the diffraction of X rays is minimized since little long-range structure or few planes exist. Thus, characterization of such materials using X rays becomes impossible. As the crystals get smaller and smaller, the XRD peaks get broader and broader and eventually are buried in the background; however, it is these “X-ray-amorphous” species that are often the most active for a given catalytic reaction.

Alternative techniques do exist, however, for obtaining information regarding the distribution and number of catalytic components dispersed within or on the support. Selective gas adsorption, referred to as chemisorption, can be used to measure the accessible catalytic component on the surface indirectly by noting the amount of gas adsorbed per unit weight of catalyst. The stoichiometry of the chemisorption process must be known in order to estimate the available catalytic surface area. One assumes that the catalytic surface area is proportional to the number of active sites and thus reaction rate. This technique has found use predominantly for supported metals. A gas that will selectively adsorb only onto the metal and not the support is used under predetermined conditions. Hydrogen and carbon monoxide are most commonly used as selective adsorbates for many supported metals. There are reports in the literature of instances in which gases such as NO and O₂ have been used to measure catalytic areas of metal oxides; however, due to difficulty in interpretation they are of limited use.

The measurements are usually carried out in a static vacuum system similar to that used for BET surface area measurements. The pressure of gas above the sample is increased and the amount adsorbed measured at equilibrium. When there is no further adsorption with increasing pressure (flat portion of Fig. 16A), it is assumed that the catalytic surface is saturated with a monolayer of adsorbate. Noting the amount of gas adsorbed and knowing its stoichiometry with the surface site, one can determine the number of catalytic sites. Approximating a value for the cross-sectional area of the catalytic component based on an assumed geometry, one can calculate its surface area and dispersion. It must be repeated that this technique measures only surface species capable of adsorbing the probing gas. The number of surface species measured is assumed proportional to active sites, but there are many cases in the literature that show no relationship between catalytic surface area and reaction rates. In such cases one evokes the concept of crystallite size effects controlling the activity or selectivity of a given reaction.

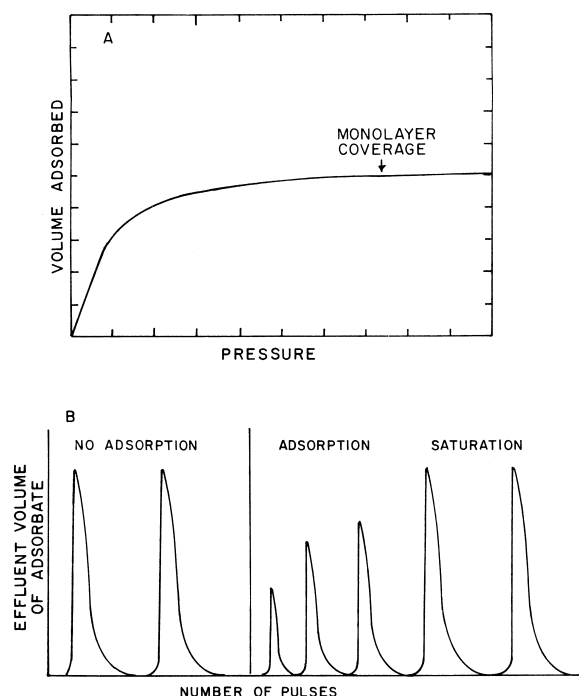


FIGURE 16 Chemisorption on a metal surface. A, Chemisorption isotherm showing approach to monolayer coverage; B, typical data from a pulsed chemisorption technique.

The static vacuum technique is traditionally used since it is an equilibrium measurement. It is time-consuming, however, and thus alternative methods exist. A dynamic pulse technique has been used over the years in which a pulse of adsorbate such as H₂ or CO is injected into a stream of inert gas and passed through a bed of catalyst. One measures the amount of gas adsorbed by comparing the amount injected with that which passes through the bed unadsorbed. As shown from left to right in Fig. 16B, the first two pulses are used for calibration and bypass the catalyst sample. The second set of pulses, passing through the catalyst, are first diminished due to adsorption. Once saturation or monolayer coverage is reached, no further adsorption from the gas phase occurs. The amount adsorbed is found by the difference in areas under the peaks compared with those under the calibration pulses. The major difference between dynamic and static methods is that the former measures only that which is strongly adsorbed, whereas the latter, performed under equilibrium conditions, measures strong and weakly chemisorbed species. Thus, static techniques usually give higher results.

2. Transmission Electron Microscopy

The above-described techniques are indirect in that they measure gas adsorption rather than the catalytic

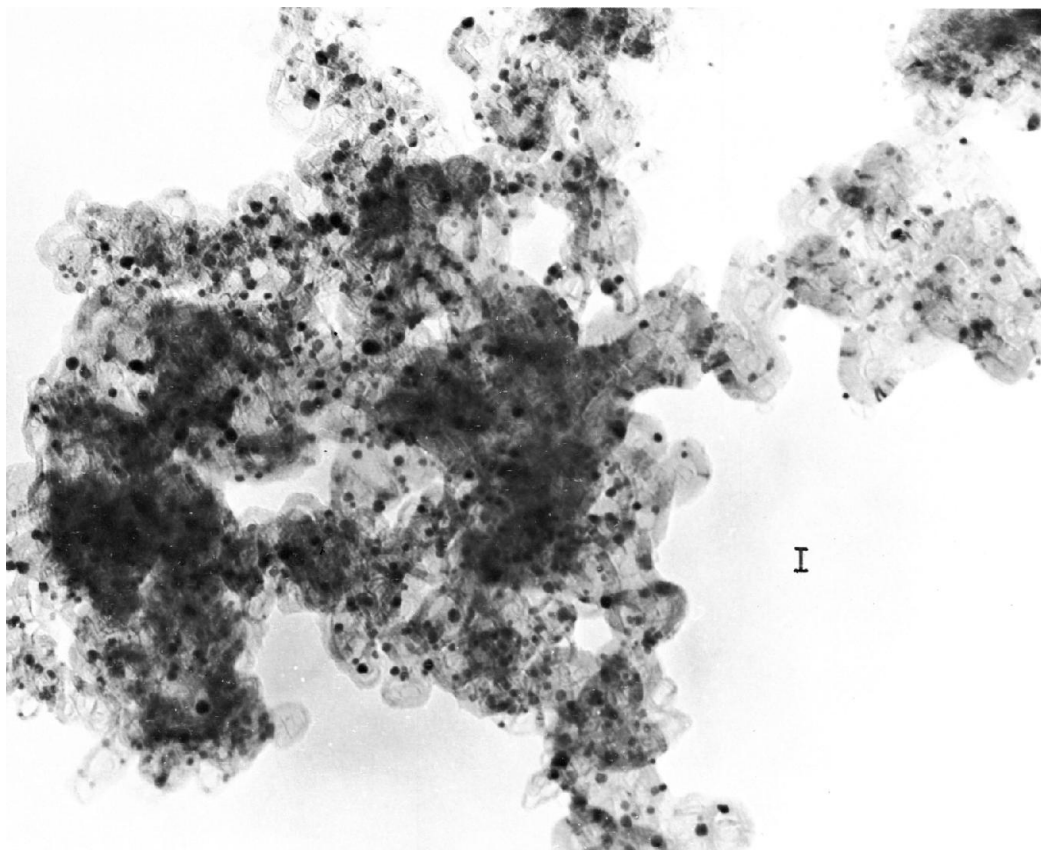


FIGURE 17 Transmission electron micrograph of platinum crystallites on a γ -alumina support. The black bar represents 100 Å.

components themselves. Transmission electron microscopy is direct in that it observes the transmission of electrons through a thin slice of material. Dense crystallites of metal or metal oxides prevent transmission and thus appear as dark spots on a photomicrograph, an example of which is shown in Fig. 17 for platinum crystallites on a γ -alumina. The sizes of the platinum crystallites range between 50 and 100 Å. By image analysis a size distribution and average crystallite size can be calculated. By assuming a shape for the crystallites the dispersion, or ratio of surface atoms to total atoms in the crystallite, can be calculated. It should be understood that this technique measures only a small fraction of the catalyst, and hence obtaining data representative of the entire sample is quite difficult. Usually many different areas must be analyzed for proper statistics to be obtained.

3. Crystallite Size by X-Ray Diffraction

The larger the crystals of a given component, the sharper are the peaks on the XRD pattern for each crystal plane. Thus, the breadth of the peak can be related to the crystal

size, which subsequently is related to the catalytic components available to reactants. The Scherrer equation relates the breadth B at half-peak-height of an XRD line due to a specific crystalline plane to the size of the crystallites l :

$$B = \frac{k\lambda}{l \cos \theta}.$$

Here λ is the X-ray wavelength, θ the diffraction angle, and k a constant usually equal to 1. As the crystallite size increases, the line breadth B decreases. Application of this technique allows the estimation of crystallite size. Figure 18 shows the crystallite size of cerium oxide growing from 50 Å in its initial state to 400 Å after thermal aging treatment.

There are problems in determining crystallite size from line broadening alone, since factors other than crystallite size contribute to the broadening, including local strain in the crystallites and shape anisotropy. Some of these problems can be overcome by the use of Fourier analysis of the peak shape. The cosine coefficients of the Fourier series can be used to determine a surface weighted average size for the crystallites.

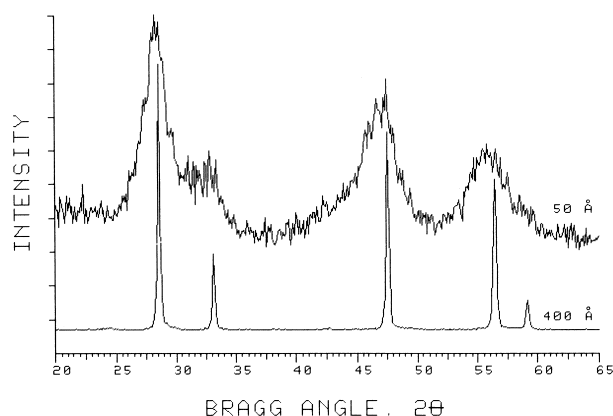


FIGURE 18 Average crystallite size measurement by X-ray line broadening. The width of characteristic X-ray lines decreases markedly as cerium dioxide powder is sintered. The crystallites grow from an initial size of 50 to 400 Å after heating in air for several hours.

Chemisorption, transmission electron microscopy, and XRD line broadening do not necessarily result in the same calculated dispersion for a given catalyst. Chemisorption may be biased toward a lower average crystallite size and line broadening toward a higher size. In fact, line broadening and chemisorption methods are not directly comparable unless Fourier analysis is applied to the X-ray data. Chemisorption and transmission electron microscope results are directly comparable.

4. Electrochemical Technique: Cyclic Voltammetry

The cyclic voltammetry technique is finding wide use in measuring catalytic surface areas for electrically conductive catalysts. Typical applications are in measuring the platinum surface area present in carbon-supported anodes and cathodes in fuel cells and the low metal surface areas found for catalytic wires, gauzes, and screens. In principle, the catalyst is immersed in an acid solution, usually H_2SO_4 , and connected to a power source. A voltage is applied and scanned cathodically, and the current resulting from reduction of H^+ to adsorbed hydrogen atoms is noted. The applied voltage is kept below that which leads to the evolution of H_2 gas. The area under the curve corresponds to the number of coulombs consumed in generating monolayer coverage. One can reverse the scan anodically, and then discharge of adsorbed hydrogen atoms to H^+ occurs. Again the area under the curve corresponds to the anodic current associated with the oxidation of a monolayer of hydrogen atoms adsorbed. The combination of current consumed (cathodic) or generated (anodic) and the time of the scan is related to the number of hydrogen atoms adsorbed as a monolayer on the metallic component, and thus the catalytic metal area is determined. Typical results for an ammonia oxidation gauze (Fig. 19), are $\sim 25 \text{ cm}^2/\text{g}$ for a fresh gauze and $\sim 250 \text{ cm}^2/\text{g}$ for a gauze that has sprouted through use. These areas are much smaller than chemisorption techniques are capable of measuring.

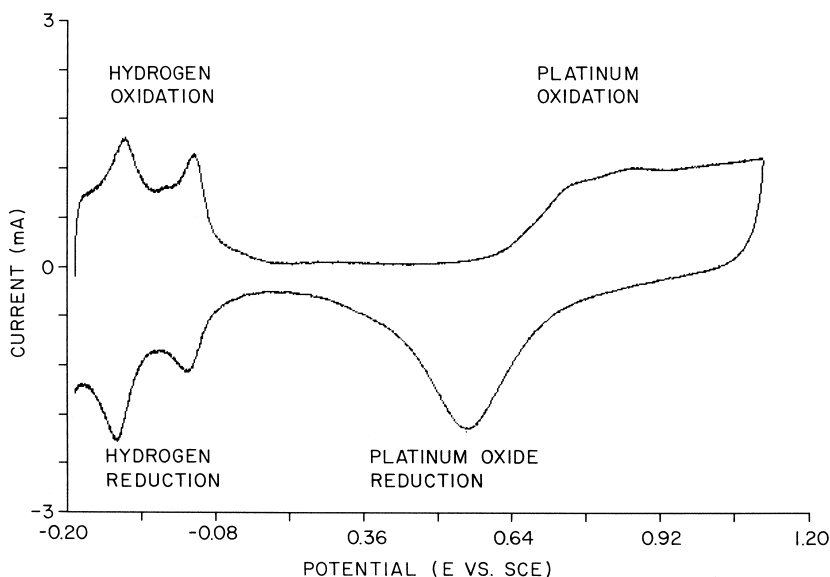


FIGURE 19 Cyclic voltammogram showing one complete anodic–cathodic sweep. The area under the hydrogen oxidation peaks is a measure of platinum surface area.

D. Surface Acidity

Many reactions are catalyzed by acid sites on the surface of the catalyst. Isomerization, polymerization, aromatization, and cracking are catalyzed by Lewis and/or Brønsted acid sites. The precise nature of these sites is open to debate; however, intuitively one can use an alkaline material to titrate acid sites and hence determine the number of such sites present. Bases, such as *n*-butylamine, with a series of Hammett indicators have been used for titrating acid sites. However, the system must be free from water contamination and the catalyst must be colorless to enable one to note indicator color changes. Diffusion of the indicators into the porous network can be very slow and require long equilibration times.

1. Gas Adsorption

Adsorption of basic gases with different alkalinities such as ammonia and/or pyridine is commonly used. The amount adsorbed is determined at a low temperature (i.e., 25°C). The strength of the sites retaining the base can be approximated by measuring the amount desorbed as a function of temperature. The stronger sites will retain the basic gas at higher temperatures than weaker sites. Thus, temperature-programmed desorption using gas chromatography or thermal analysis techniques is suitable. In Fig. 20 the amount of ammonia desorbed from three $\text{La}_2\text{O}_3/\text{Cr}_2\text{O}_3/\text{Al}_2\text{O}_3$ supports prepared differently is shown as a function of desorption temperature. As the temperature is increased, the gas adsorbed on the weaker sites desorbs. Curve A shows significant retention at 400°C, indicating that strong acid sites are present; the material

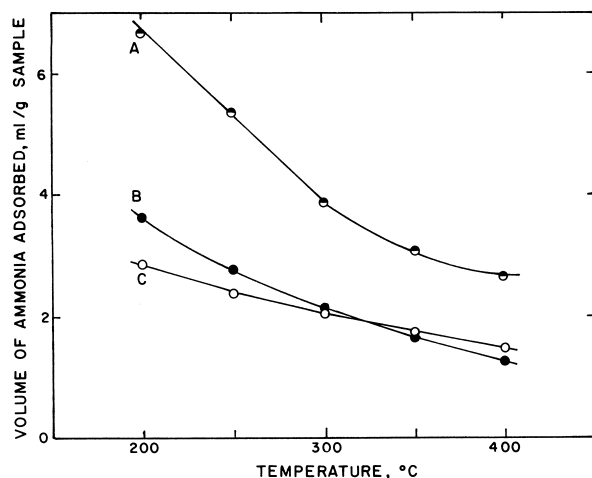


FIGURE 20 Amount of NH_3 adsorbed on acid sites measured by temperature-programmed desorption. The three curves show different acid site distributions for the same catalyst composition prepared by different methods.

represented by curve B has fewer total acid sites and fewer strong sites, as indicated by the small amount of retained base at 400°C. Sample C has the fewest total acid sites but more strong sites than B.

The strength of acid sites can be approximated by determining the heat of adsorption using the Van't Hoff isochore. In this method the pressures P required to give a fixed surface coverage at a series of temperatures T is determined. The derivative of $\ln P$ with respect to T is related to the isosteric heat of adsorption. This technique is not used routinely. Nor is that of IR spectroscopy, where it is reported that Lewis and Brønsted sites can be differentiated by using substituted compounds.

2. Infrared Analysis

Infrared absorption can be used to estimate the relative amounts of Lewis and Brønsted acid sites on cracking catalysts. Bases complex with Lewis acid sites while proton transfer to the base occurs at Brønsted acid sites. Each has distinct, well-resolved infrared bands. For example, pyridine forms a complex with the Lewis acid site and produces an infrared absorption band at approximately 1450 cm^{-1} . Pyridinium ions form at Brønsted sites and produce an absorption band at approximately 1540 cm^{-1} . The relative intensities of these two bands can be used to estimate the relative amounts of Lewis vs. Brønsted acid sites.

The technique is not without its problems. One criticism of the model is that pyridine adsorbed on Lewis acid sites could hydrogen bond with nearby surface hydroxyls. Then proton transfer could take place, and in the infrared spectrum the pyridine would appear to have adsorbed on a Brønsted acid site. Also, attempts to measure acid strengths by following infrared spectral changes during desorption of the base can be complicated by decomposition of the base. Decomposition of the pyridinium ion at a given temperature may suggest a Brønsted acid site with a given acid strength while, in fact, the Brønsted acid site remains intact and does not decompose until a much higher temperature. Measurement and interpretation of surface acidity remains a controversial subject.

E. The Surface of Catalysts

1. Surface Composition

Such techniques as XRD and electron microscopy measure the structure and/or chemical composition of catalysts extending below the surface where reaction occurs. The composition of the surface is usually different from that of the bulk, and thus its analysis must be carried out by techniques specific to the surface. It is on these surfaces that

the active sites exist and where chemisorption, chemical reaction, and desorption take place.

Most commonly a freshly reduced metal or metal-supported catalyst will have a layer of oxygen, either strongly chemisorbed or fully oxidized, on its surface. Because of this it is common practice in hydrogenation reactions to pretreat the catalyst in a stream of H_2 in the reactor to remove all traces of surface oxygen before performing the catalytic reactions. Also, throughout the years catalytic scientists have dealt with the problem of pyrophoric metals by intentionally blanketing the catalyst with a carefully controlled amount of O_2 or CO_2 after reduction to prevent bulk oxidation. These protective layers are removed by reduction and/or heat treatment in order to permit catalysis to occur.

Platinum–rhodium alloys in the form of wires woven into screens or gauzes are catalysts for the production of hydrogen cyanide by the reaction of ammonia, methane, and oxygen. The surface of these wires may contain impurities from the manufacturing operations, and thus surface analysis as opposed to bulk analysis is more important for predicting reaction rates. During ammonia oxidation, the platinum–rhodium alloy becomes surface enriched with less active rhodium due to the volatility of the platinum oxide, causing the reaction rate to diminish gradually.

Sulfur compounds adsorb onto surface-active metal (or metal oxide) sites, causing deactivation in a large number of petroleum, petrochemical, and chemical catalytic applications. Acidic catalysts such as zeolites and promoted aluminas are poisoned by nitrogen compounds by chemisorption onto active sites also located on the surface.

2. Surface Techniques

The tools available for surface composition characterization are electron spectroscopy for chemical analysis (ESCA), Auger spectroscopy (AES), ion scattering spectroscopy (ISS), and secondary ion mass spectroscopy (SIMS). ESCA spectroscopy is used more widely than the others for studying the surface composition and oxidation states of industrial catalysts, and thus its application will be discussed in limited detail.

The acronym ESCA refers to the technique of bombarding the surface with X-ray photons, which produce the emission of characteristic electrons measured as a function of electron energy. Because of the low energy of the characteristic electrons, the depth to which the analysis is made is only ~ 20 Å. The composition of this thin layer as a function of depth can be determined by sputtering away layers of the surface and analyzing the underlying surfaces. A number of important catalytic properties have been studied by this technique, including oxidation state of the active species, interaction of a metal with an

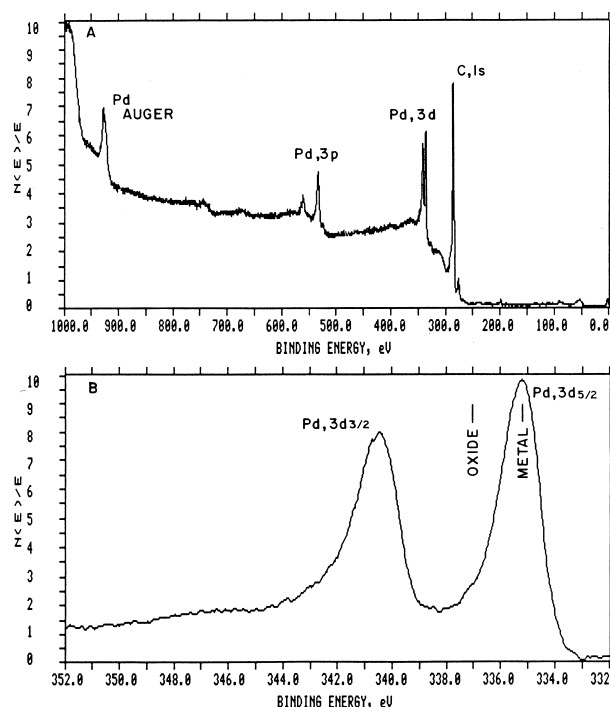


FIGURE 21 ESCA spectrum of a palladium-on-carbon catalyst. A, Survey scan locating the carbon 1s and the palladium 3p and 3d peaks; B, scale expansion around the palladium 3d doublet. The small shoulder at the oxide location is assigned to a chemisorbed oxygen state.

oxide support, changes in oxidation state on activation of the catalyst, and deactivation of the catalyst by poisons.

Figure 21 provides an example of the use of ESCA to define an oxidation state of a freshly reduced palladium-on-carbon hydrogenation catalyst exposed to the air. The metallic palladium peaks (Fig. 21a) are quite evident, indicating no bulk oxidation occurred. There is a strong peak for carbon, probably due to adsorbed CO_2 from the air. The presence of a small amount of PdO is suggested at 337 eV in Fig. 21B. This peak is a shoulder on the palladium $3d_{5/2}$ peak and most likely represents a surface layer of oxide on the palladium. This information could not be conveniently obtained by XRD because small palladium (or PdO) crystallites cannot diffract X rays. Furthermore, XRD measures bulk properties and would not “see” surface oxides even if the crystallite sizes were sufficiently large to be XRD sensitive. We can therefore expect to see more frequent use of ESCA or other surface sensitive techniques to monitor the surface of catalytic materials.

IV. COMPLEMENTARY TECHNIQUES

Characterization techniques described in this section are used primarily to support the more routine methods used

for measuring the chemical and physical properties of a catalyst. The assignment of these techniques to this category has been somewhat arbitrary. Some of the techniques may become routine as instrumentation and the treatment of data improve, whereas others will provide fundamental insight into the mechanisms of catalytic reactions at a solid surface. All of these techniques are powerful enough to provide information on the physical and chemical properties of a catalyst, such as acid sites, oxidation state of a supported metal, micropore structure, and other important characteristics not easily ascertained by less sophisticated methods.

A. Spectroscopic Techniques

The spectroscopic techniques described in this section include IR, Raman, and UV-visible spectroscopy, nuclear magnetic resonance (NMR) and electron spin resonance (ESR) spectroscopy, and extended X-ray absorption fine structure (EXAFS) spectroscopy. Techniques based on particle scattering, transitions in the nucleus, and radioisotope techniques that produce radiation that is a measure of the chemical environment are described in Sections IV.B and C. Some of these techniques, such as IR and UV-visible spectroscopy, have been applied to studies of catalysts for more than 30 years, whereas others, such as EXAFS, are relatively new to catalytic studies.

1. Infrared Spectroscopy

Infrared spectra of a catalyst can be obtained by either of two techniques: transmission through the sample or reflection from its surface. The transmission technique is by far the more widely used of the two and is applicable generally to oxides and supported metals. Most of the oxides have been porous glass, very fine powders made by high-temperature hydrolysis of silicon or aluminum halides, and aerogels of silica, alumina, and mixed oxides. A number of investigations have been directed toward the study of hydroxyl groups on high surface area oxides and their behavior toward chemisorption of various molecules. However, most of the studies have involved metals supported on these oxides. Infrared spectra have been obtained on a number of molecules chemisorbed on supported metals, including water, ammonia, carbon monoxide, and a variety of low molecular weight hydrocarbons. There are disadvantages, in that the samples are small and it is often difficult to eliminate contamination. Heterogeneity of the surface may cause difficulties in interpretation of the data, and the oxide support for the metal usually has a cutoff frequency such that the entire spectrum is not available below frequencies of 1000 to 1800 cm^{-1} . Some of the problems associated with using these oxides as supports for the metals have been overcome by evaporating

metal films onto transparent substrates, such as calcium fluoride and sodium chloride. Thin evaporated films suffer from a loss of sensitivity. Modern Fourier transform instrumentation has eliminated some of the problems of sensitivity and sample preparation that plagued the dispersive instruments.

The chemisorption of CO on supported metals such as platinum and palladium illustrates the technique. Figure 22 shows the development of spectra of chemisorbed CO on Pd/SiO₂ as surface coverage increases to a monolayer. Spectrum a, taken after exposure of the catalyst to 131 Torr of CO and then evacuation of the cell at room temperature, shows a strong band at 1979 cm^{-1} and a small band at 2050 cm^{-1} assigned to bridged and linear bonded CO, respectively. As additional CO chemisorbs, the 1979 cm^{-1} band is converted to a band at 1995 cm^{-1} for the bridged species and a band at 2103 cm^{-1} increases for the linear species. These spectra clearly show a changing surface stoichiometry from bridged to linear species as CO chemisorption approaches monolayer coverage.

Both the relative intensities of these two bands and the frequency shifts are highly dependent on the conditions

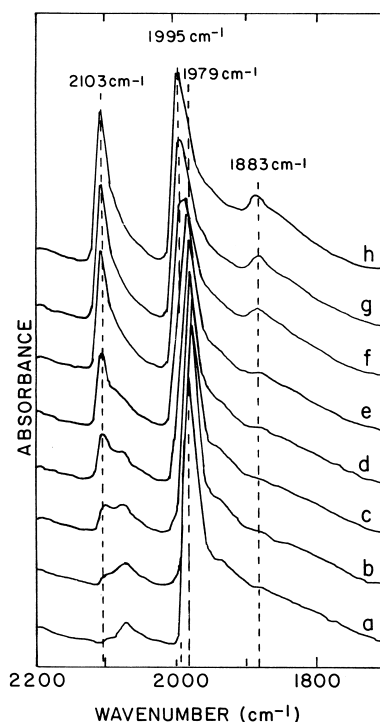


FIGURE 22 Infrared spectra of CO adsorption on Pd/SiO₂ as a function of surface coverage. (a) Follows saturation CO coverage at 300 K, evacuation for 30 min, and cooling to 80 K. (b–h) Incremental addition of CO to the cooled Pd/SiO₂. Reprinted with permission from Gelin, P., Siedle, A. R., and Yates, Jr., J. T. (1984). *J. Phys. Chem.* **88**, 2978–2985. Copyright 1984, American Chemical Society.

of the metal surface, the nature of the metal, interaction of the metal with the support material, and other features of heterogeneity, such as variations in crystallite size and surface coverage. A completely satisfactory explanation of CO chemisorption on metals from IR data has yet to be realized. Changing ratios of linear to bridged species signals caution in interpreting the dispersion of metal on a support from CO chemisorption measurements. Dispersion calculations from chemisorption data must assume a surface stoichiometry for the chemisorbed molecule, and IR results have demonstrated that an assumed model can easily be off by 20 or 30% depending on the ratio of linear to bridged species.

The reflection technique has not been used as extensively as transmission. Its slow development may be attributed to several factors. It is used primarily on highly polished metal surfaces including those involved in fundamental studies of single crystals. The theoretical framework for reflection IR spectra has been developed only recently. Ultrahigh-vacuum techniques are required and modifications are needed for standard IR spectrometers. Since the reflection technique can be used with single crystals of metals, it is a bridge between the more sophisticated surface techniques used in surface science and the IR studies of the more practical catalysts.

Recent studies using the reflection technique have revealed a flaw in interpretation of the data from transmission studies of CO chemisorption on metals. Studies of CO chemisorption on polycrystalline copper and several different faces of single crystals of copper have shown marked variations in the frequency shift of the IR spectra. Similar results have been obtained on palladium films. The polycrystalline and high-index faces show an absorption band at $\sim 2100\text{ cm}^{-1}$, whereas low-index faces show a corresponding band at $\sim 2070\text{ cm}^{-1}$. This suggests that the support may influence the development of high-index planes and steps on the surfaces of small crystallite deposits. On palladium films deposited at low temperatures a similar CO absorption band is at 2112 cm^{-1} . Films annealed at elevated temperatures, 150°C , have absorption bands in a region between 1930 and 1990 cm^{-1} . These results suggest that the original assignment of linear and bridged CO species on supported palladium may, in fact, be associated with low- and high-index faces on the crystallites. The concept of the support strongly influencing the type of crystallite faces formed on impregnation and deposition of metal on oxide surfaces is reinforced by these observations.

Some recent developments in IR techniques have included IR photoacoustic and photothermal beam deflection spectroscopy. In photoacoustic spectroscopy the IR wave incident to the solid surface of the catalyst is absorbed by the sample. The radiation is converted to a

thermal wave that acts as a piston on an inert gas layer at the surface of the sample and produces a sound wave that is detected by a microphone. Although the application is relatively new to catalysis, it has been used successfully in the study of acid sites on silica-alumina catalysts and similar materials. The photothermal deflection technique is an outgrowth of the photoacoustic method. In this technique the IR radiation is focused on the surface of the catalyst sample, producing thermal gradients in the gas just above the sample. A laser beam probe passes through the gas just above the sample and is deflected by the changes in the gas density. The deflected beam is detected with a position sensitive detector, and the angle is measured between the deflected and undeflected beams. The photothermal technique has the distinct advantage of being adaptable to *in situ* measurements of catalyst behavior. The application of both techniques to catalyst studies is new and relatively unexplored.

2. Raman Spectroscopy

Application of Raman spectroscopy to a study of catalyst surfaces is increasing. Until recently, this technique had been limited to observing distortions in adsorbed organic molecules by the appearance of forbidden Raman bands and giant Raman effects of silver surfaces with chemisorbed species. However, the development of laser Raman instrumentation and modern computerization techniques for control and data reduction have expanded these applications to studies of acid sites and oxide structures. For example: The oxidation-reduction cycle occurring in bismuth molybdate catalysts for oxidation of ammonia and propylene to acrylonitrile has been studied *in situ* by this technique. And new and valuable information on the interaction of oxides, such as tungsten oxide and cerium oxide, with the surface of an alumina support, has been obtained.

Future applications of the Raman technique to catalyst studies appear promising.

3. Magnetic Resonance

Magnetic resonance spectra are obtained from species with net magnetic moments. Two species are recognized: a rotating electron that produces electron paramagnetic resonance and a spinning nucleus with a net magnetic moment that produces nuclear magnetic resonance.

Resonance spectra are obtained by placing a sample in a magnetic field. The magnetic field removes the degeneracy of magnetic states of the system. The addition of a small oscillating field perpendicular to the direction of the magnetic field adds an additional small increment of energy to the system. When the frequency of the

oscillating field matches the energy difference between magnetic states, resonance occurs and energy is absorbed from the oscillator.

Electron paramagnetic resonance (EPR) has been used extensively in studies of the mechanism of catalytic reactions. It has been used to identify free radicals and ion-radicals formed by chemisorbed species on catalytically active sites and to study the structure and distribution of paramagnetic catalytic sites such as those produced by transition metals or metal ions on a catalyst surface. EPR remains primarily a research tool for studying mechanisms of catalytic reactions.

The principles of NMR are similar to those of EPR. In this case the nucleus is the probe. The chemical information is contained in shifts in the nuclear magnetic energy levels as a result of coupling with the external electronic system of the atom. Until recently very few examples of applications of NMR to catalysts appeared in the literature. Before the recent discovery of the phenomenon of magic angle spinning, solids exhibited broad, featureless absorption bands in the NMR region. Magic angle spinning removes the dipolar broadening found in the solid state, and sharp absorption lines result. The application of magic angle spinning to catalyst characterization has not been fully developed and this technique could become an important characterization tool in the future.

4. EXAFS and XANES

Extended X-ray absorption fine structure (EXAFS) and X-ray absorption near edge fine structure (XANES) have been used with great success to characterize highly dispersed supported metals. EXAFS can provide information on the local structure of highly dispersed metal on a support, such as alumina, and XANES, measured at the same time, provides information on the valence state of the metal.

With EXAFS oscillations are observed in the photon energy spectrum just above the X-ray absorption edge of an element. Theoretical studies of these oscillations have demonstrated that they contain structural information on the immediate surroundings of the atom undergoing X-ray absorption. When an X-ray photon is absorbed by an atom, a photoelectron is emitted. The outgoing photoelectron wave may be reflected from the neighboring atoms. The constructive and destructive interference between the outgoing and reflected photoelectron waves produces the oscillations observed just above the X-ray absorption edge. The structural information is extracted by a Fourier transform of the oscillations. The resulting absorption spectrum shows peak intensity as a function of nearest-neighbor distances. The abscissa is calibrated from the oscillation

spectrum of a reference bulk material with known nearest neighbor distances. The precision of the measurements of distances to neighboring atoms is high, and these distances can be assigned to bond type, such as metal-metal or metal-oxygen. The number of nearest and next nearest neighboring atoms is obtained from the intensity of the absorption peaks.

Most of the published EXAFS studies have been concerned with the structure of very small metal crystallites supported on high surface area materials such as silica gel, alumina, or zeolites. Typically, crystallites are smaller than 100 Å in diameter.

As an example of the application of EXAFS to catalyst systems, consider studies of platinum crystallites supported on silica gel, alumina, and Y zeolite. In each case the coordinate number for the nearest neighbor atoms was smaller than in the bulk metal. It was 8 for platinum on silica gel and 7.2 for platinum on alumina. The nearest neighbor distances were similar to bulk distances, namely, 2.775 Å for platinum on silica gel and Y zeolites. However, the distance was much smaller for platinum on alumina. The results suggest that small crystallites, similar to bulk metal, are formed on silica and Y zeolite supports, but strong interactions exist between platinum and alumina to produce the small interatom distances and low coordination number. The absorption near edge fine structure results from electron transitions to bound states in the adsorbing species. The fine structure is related to the electronic structure of the adsorbing atom and provides information on its valence state and interactions with the support or with chemisorbed species. ZANES' data on the platinum catalyst mentioned above demonstrated that in the reduced state the supported platinum was electron deficient compared to the bulk metal, indicating that the platinum 5*d* electrons form bonds with the support.

Unfortunately, EXAFS cannot be routinely applied to the characterization of catalysts. Practical applications require high-intensity X-ray sources obtainable only at synchrotron installations.

B. Particle Diffraction and Scattering

The scattering of neutrons, ions, or electrons results in changes in both the energy and momentum of these particles. Structural information is obtained from the measurement of these changes.

1. Neutrons

Structural information important for catalyst characterization can be obtained from neutron diffraction, inelastic scattering, and small-angle scattering. Each experimental technique yields a different type of structural information.

The measurement of small-angle scattering of neutrons provides information similar to that furnished by other small-angle techniques, such as small-angle X-ray diffraction. Small-angle scattering results from the differences in neutron scattering length density of pores and particles in the solid matrix of the catalyst. The method does not distinguish between open and closed pores. A combination of the two techniques provides more complete characterization of a supported heavy-metal catalyst. A combination of neutron and X-ray scattering and gas adsorption on a supported metal catalyst measures pore size, metal particle size, and whether metal has been encapsulated in small pores by the conditions of its usage.

Inelastic scattering of neutrons yields neutron scattering spectra that measure the vibrational energy levels of the material under study. For example, the chemisorption of water on Raney nickel was shown by inelastic scattering both to produce hydroxyl groups and to chemisorb water molecules on the surface at less than monolayer coverages.

2. Ion Diffraction

Ion scattering spectrometry and secondary ion mass spectrometry are the best-known types of ion-beam-induced analyses applicable to catalysis. Other ion-beam techniques have not enjoyed wide use, probably because accelerators are required to produce sufficiently energetic ion beams.

The ISS and SIMS techniques are sensitive to less than monolayer coverages and detect many of the lighter elements, including hydrogen. Their application to practical catalysts is somewhat limited. The sampling area is $\sim 1 \text{ mm}^2$, and the results obtained from examining the surface of a typical catalyst pellet can be very ambiguous.

Other ion-beam techniques applicable to catalysts are proton-induced X-ray emission (PIXE), Rutherford backscattering, and resonance ion-beam backscattering. The proton-induced X-ray emission technique is very similar to electron microprobe analysis. Characteristic X-rays are generated by bombardment of the sample with a proton beam. The advantages of PIXE are high sensitivity, particularly for light elements, and a high signal-to-noise ratio, which allows analysis even of trace elements in a few minutes.

In Rutherford backscattering the sample is bombarded with light ions such as helium ions, and the energy of the light ions that have been backscattered from the sample is measured. When the ion is reflected back from the surface layer the energy lost is inversely proportional to the atomic number of the scattering species. By measuring the energy loss the element can be identified and a simple spectrum obtained. The broadening of the resonant

backscatter line has been used to measure pore sizes in porous materials. The advantage of the technique is that it measures smaller pore sizes than can be measured by gas adsorption techniques, and it is capable of measuring pore size distributions as a function of depth within the pellet by varying the energy of the incoming ion beam.

C. Radioisotope Techniques

This section deals with spectroscopic methods that depend on nuclear events. Radiotracer techniques that are very valuable in studying the mechanisms of catalytic reactions have been omitted.

1. Mössbauer Spectroscopy

Mössbauer spectroscopy depends on low-level nuclear transitions that emit or absorb low-energy γ rays. Coupling of the nucleus with its electronic surroundings results in changes in the spectrum of the nuclear energy levels. These changes can be interpreted in terms of the valence state of the Mössbauer atom and the electronic and magnetic surroundings of its environment. The three basic parameters observed in Mössbauer spectra are the isomer shift δ , quadrupole splitting ΔE_q , and magnetic splitting.

The isomer shift is a result of changes in the electron density at the nucleus, which produce small changes in the nuclear energy levels. The result is shifts in the centroid of the Mössbauer spectrum relative to a standard material. Asymmetry in the electric field surrounding the nucleus produces splitting of the energy levels. Additional splitting results from a permanent magnetic field surrounding the nucleus. The benefits of the detailed information on oxidation state at the atomic level is offset by the limited number of elements with useful Mössbauer isotopes. The application of the ^{57}Fe isotope has been widespread not only in the characterization of iron catalysts, but also as an atomic probe in studies of other supported metal catalysts.

The spectra in Fig. 23 show the effect of the chemisorption of NH_3 on the ^{57}Fe resonance of a highly dispersed Fe/SiO_2 catalyst. The iron is in the ferrous state following the hydrogen reduction of microcrystalline ferric oxide deposits. The addition of NH_3 to the reduced and outgassed sample has a marked effect on peak 2 of the spectra. Peak 2 decreases in relative area as small doses of NH_3 are added. Since peak 2 has been assigned as half of a doublet produced by surface ferrous ions, the relative area of this peak is used as a measure of surface sites available for chemisorption. There is no noticeable change in the spectrum with the addition of the first small amounts of adsorbate. However, the relative area under peak 2 begins to decrease with further addition of NH_3 , starting at $\sim 2.0 \times$

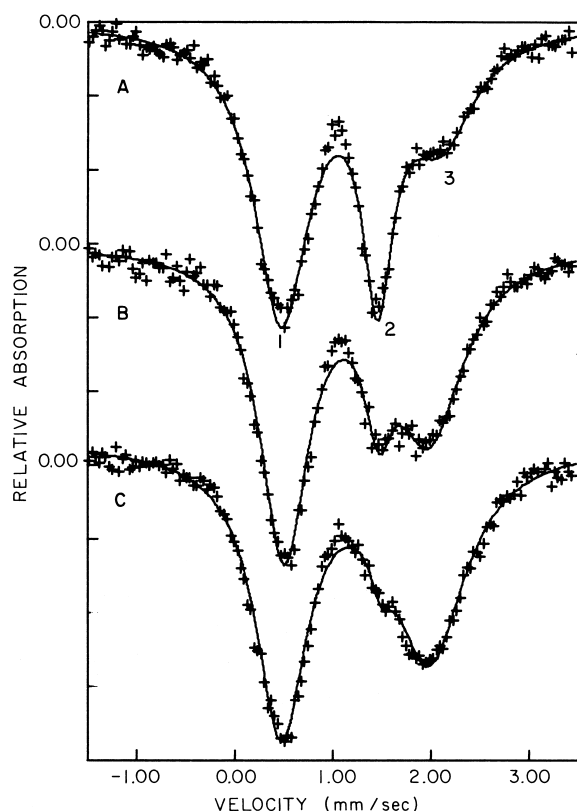


FIGURE 23 Effect of ammonia chemisorption on the Mössbauer spectrum of a highly dispersed iron on silica gel in its ferrous state. (A) 0.5×10^{-2} mmol NH_3 ; (B) 2.99×10^{-2} mmol NH_3 ; (C) 4.31×10^{-2} mmol NH_3 . Adapted from Hobson, Jr., M. C., and Gager, H. M. (1970). *J. Colloid Interface Sci.* **34**, 357–364, by permission of Academic Press, Inc.

10^{-2} mmol of added adsorbate. The decrease in the area is rapid at first and then slowly levels off to a minimum value of 0.02 at $\sim 5.0 \times 10^{-2}$ mmol of added adsorbate.

Desorption of NH_3 at room temperature does not cause any changes in the spectrum. On raising the temperature of the sample to 100°C , NH_3 begins to desorb and peak 2 begins to increase in size. Increments of NH_3 continue to desorb as the temperature is raised in steps until the original spectrum is recovered at a desorption temperature of 300°C . At this temperature, the NH_3 mass balance is not complete. Some of the NH_3 remains strongly bound to surface sites on the silica gel. The amount of NH_3 that forms a monolayer coverage of the ferrous ion surface sites is $\sim 3 \times 10^{-2}$ mmol, or just enough to form a 1:1 surface complex with the ferrous species.

Since the area under peak 2 does not begin to decrease until a small amount of NH_3 has been added, and fully recovers before all of the NH_3 has been desorbed, the initial increment of NH_3 must be strongly adsorbed on silica sites. Similar results have been obtained in IR studies of

NH_3 desorption on porous glass. The addition of NH_3 had no effect on the IR band of surface silanol groups until enough NH_3 had been added to cover $\sim 1\%$ of the surface. Thus, the Mössbauer and IR results provide additional evidence for the complexity of surface acid site measurement and interpretation.

2. Positron Spectroscopy

Like Mössbauer spectroscopy, positron spectroscopy is a nuclear process for probing the chemical and physical environments of solid materials. It is a more versatile technique than Mössbauer spectroscopy, but it does not provide as much information on the structure surrounding the nucleus or the immediate environment of the atom.

Applications of the technique to heterogeneous catalysts have been few, but they have demonstrated that the method is useful for catalyst characterization. For example, the lifetime of the orthopositronium species is inversely proportional to the number of Brønsted acid sites present in alumina–silica cracking catalysts. This interpretation was derived from a correlation between the activity for the alkylation of cumene and the lifetime of the orthopositronium species.

As the time resolution of the equipment improves and computer programs become available for data reduction, additional applications of the technique to heterogeneous catalysis can be expected.

SEE ALSO THE FOLLOWING ARTICLES

ADSORPTION (CHEMICAL ENGINEERING) • BATCH PROCESSING • CATALYSIS, HOMOGENEOUS • CATALYSIS, INDUSTRIAL • ELECTROCHEMISTRY • INFRARED SPECTROSCOPY • MÖSSBAUER SPECTROSCOPY • NUCLEAR MAGNETIC RESONANCE • RAMAN SPECTROSCOPY • SCANNING ELECTRON MICROSCOPY • SURFACE CHEMISTRY

BIBLIOGRAPHY

- Anderson, R. B., and Dawson, P. T. (eds.) (1976). "Experimental Methods in Catalytic Research," Academic Press, New York.
- American Society for Testing and Materials (1984). "Standards on Catalysis," 2nd Ed., ASTM, Philadelphia.
- Bond, G. C., and Webb, G., eds. (1984). "Catalysis," Vol. 6, Royal Soc. Chem., London.
- Delannay, F. (1984). "Characterization of Heterogeneous Catalysts," 1st Ed., Dekker, New York.
- Deviney, M. L., and Gland, J. L., eds. (1985). "Catalyst Characterization Science: Surface and Solid State Chemistry," ACS Symp. Series No. 288, American Chemical Society, Washington, D.C.

- Eley, D. D., Pines, H., and Weisz, P. B. (eds.) (1985). "Advances in Catalysis," Vol. 33, Academic Press, New York.
- Heinemann, H., and Carberry, J. J. (eds.) (1984). "Catalysis Reviews—Science and Engineering," Vol. 26, Dekker, New York.
- Kaye, B. H. (1981). "Direct Characterization of Fine Particles," Wiley, New York.
- Satterfield, C. N. (1980). "Heterogeneous Catalysis in Practice," 1st Ed., McGraw-Hill, New York.
- Thomas, J. M., and Lambert, (eds.) (1980). "Characterization of Catalysts," Wiley, New York.
- Whyte, T. E., Jr., Dalla Betta, R. A., Derouane, E. G., and Baker, R. T. K., (eds.) (1984). "Catalytic Materials: Relationship Between Structure and Reactivity," ACS Symp. Ser. No. 248. Am. Chem. Soc., Washington, D.C.



Chemical Process Design, Simulation, Optimization, and Operation

B. Wayne Bequette

Rensselaer Polytechnic Institute

Louis P. Russo

ExxonMobil Chemical

- I. Background
- II. Process Models
- III. Process Simulation
- IV. Optimization
- V. Process Design
- VI. Design for Operability and Flexibility
- VII. Process Operation

GLOSSARY

Control Using a manipulated input variable (often a valve regulating a stream flowrate) to maintain a measured output variable (often temperature, level, pressure, or composition) at a desired value (setpoint).

Flowsheet A schematic representation of the chemical process streams and unit operations involved in a chemical process.

Operability The ability of a process to be regulated at desired operating conditions under uncertainty in feed or utility streams.

Optimization Mathematical procedures to enable the best selection from a set of many options, for example, in the context of process design, the selection of

equipment to minimize the annualized costs (including operating costs and capital investment amortized over several years).

Process design The specification of process equipment and flowstreams necessary to produce chemical processes to meet desired production specifications.

Process model A mathematical description of a chemical process that, if solved, enables prediction of the behavior of the chemical process.

Simulation The numerical solution of a mathematical model, often refers to the numerical solution of a model of an entire process flowsheet.

Unit operation Typically, an equipment item that performs a chemical/physical transformation of a process stream, for example, a heat exchanger may increase the

temperature of a process feedstream by using steam as a heating media.

I. BACKGROUND

The field of process systems engineering refers to the various techniques to design, simulate, optimize, and operate chemical processes. A majority of chemical processes operate in a continuous fashion; that is, raw material is continuously fed to the manufacturing process and product is continuously produced. Large manufacturing processes often run for 18 months or more without any major shut-down. For this reason, most process systems engineering techniques have been applied to continuous processes, which are the focus of this article. Batch processes are presented elsewhere.

Many consumer products are produced, at least in part, using chemical processes. A characteristic chemical process involves a chemical and/or physical transformation of raw materials into products or intermediates that are then further processed. Process flowsheets or process flow diagrams are used by process engineers to depict the flow of process streams through the basic unit operations involved in a chemical manufacturing process. A unit operation typically refers to a vessel where a chemical or physical transformation occurs. Examples include chemical reactors and distillation columns.

A simplified process flow diagram of a classical chemical process is shown in Fig. 1. Here, the feedstream is mixed with a recycle stream and passes through a heat exchanger into a chemical reactor. The reactor effluent is processed by the separations device into two streams—one is the product stream and the other is a recycle stream which often contains a significant amount of the original feed component.

The chemical process industries include the following:

- Petroleum refining and petrochemicals
- Food and beverages
- Pulp and paper

- Plastics and polymeric materials
- Air products (oxygen, nitrogen, etc.)
- Consumer products (detergents, etc.)
- Agricultural chemicals
- Pharmaceuticals
- Inorganic chemicals (sodium hydroxide, etc.)
- Textiles
- Minerals processing
- Biotechnology
- Electronic materials and semiconductor device manufacturing

In this article, we provide an overview of a number of techniques used by process engineers. We will provide references to numerous sources for details of the methodologies or applications.

II. PROCESS MODELS

In order to make design or operation decisions a process engineer uses a process model. A process model is a set of mathematical equations that allows one to predict the behavior of a chemical process system. Mathematical models can be fundamental, empirical, or (more often) a combination of the two. Fundamental models are based on known physical–chemical relationships, such as the conservation of mass and energy, as well as thermodynamic (phase equilibria, etc.) and transport phenomena and reaction kinetics. An empirical model is often a simple regression of dependent variables as a function of independent variables. In this section, we focus on the development of process models, while Section III focuses on their numerical solution.

Models are either dynamic or steady-state. Steady-state models are most often used to study continuous processes, particularly at the design stage. Dynamic models, which capture time-dependent behavior, are used for batch process design and for control system design. Another classification of models is in terms of lumped parameter or distributed parameter systems. A lumped parameter system

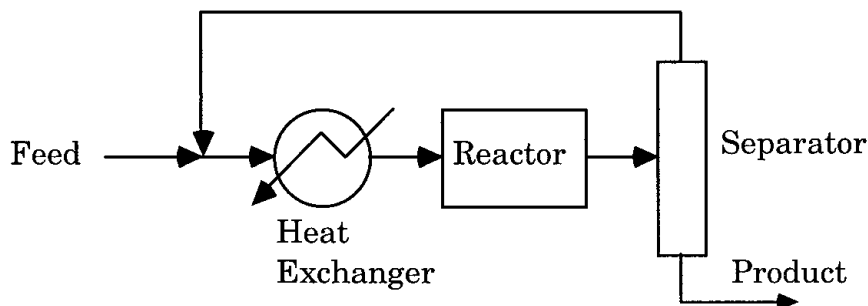


FIGURE 1 Flow diagram of a simple chemical process with recycle.

assumes that a variable of interest (temperature, for example) changes with only one independent variable (time, but not space, for a dynamic systems example). A typical lumped parameter system is a perfectly mixed (stirred) tank, where the temperature throughout the tank is uniform (there is no spatial gradient). A distributed parameter system has more than one independent variable; for example, temperature may vary with both spatial position and time.

A. Lumped Parameter Models

Consider the dynamic behavior of a process that can be considered perfectly mixed. The lumped parameter model has the following form:

$$\dot{x} = \frac{dx}{dt} = f(x, p, u) \quad (1)$$

where x = states, p = parameters, u = inputs, t = time.

To illustrate the basic principles, consider a perfectly mixed, isothermal chemical reactor, with a series reaction of the form:



Assuming constant density and volume, the following modeling equations can be written, where it is assumed that each reaction is a first-order decomposition:

$$\begin{aligned} \frac{dC_A}{dt} &= \frac{F}{V}(C_{Ain} - C_A) - k_1 C_A \\ \frac{dC_B}{dt} &= \frac{F}{V}(C_{Bin} - C_B) + k_1 C_A - k_2 C_B \\ \frac{dC_C}{dt} &= \frac{F}{V}(C_{Cin} - C_C) + k_2 C_B \end{aligned} \quad (2)$$

where F = volumetric flowrate; V = reactor volume; C_A , C_B , C_C = concentrations of components A , B , and C , respectively; k_1 , k_2 = reaction rate constants; and C_{Ain} , C_{Bin} , C_{Cin} = feedstream concentrations of components A , B , and C , respectively.

In terms of the state variable notation of Eq. (1), there are three states, three parameters, and four inputs (often the feedstream compositions of components B and C will be zero):

$$\begin{aligned} x &= [C_A \quad C_B \quad C_C]^T \\ p &= [k_1 \quad k_2 \quad V]^T \\ u &= [F \quad C_{Ain} \quad C_{Bin} \quad C_{Cin}]^T \end{aligned}$$

Notice that steady-state models can be obtained by setting the derivative (accumulation) terms to zero. In this case, the ordinary differential equations become algebraic equations. For a fixed set of values of the parameters and inputs, the algebraic equations have the form:

$$f(x) = 0$$

which can be solved using numerical techniques, as shown in Section III.

B. Distributed Parameter Models

Consider a tubular reactor where a chemical reaction changes the concentration of the fluid as it moves down the tube. Assuming first-order chemical reaction, isothermal reactor, and constant density, the modeling equation is

$$\frac{\partial C_A}{\partial t} = -v_z \frac{\partial C_A}{\partial z} + D_{Az} \frac{\partial^2 C_A}{\partial z^2} - k C_A \quad (3)$$

where z is the spatial coordinate, C_A is the concentration of component A , D_{Az} is the diffusion coefficient, k is the reaction rate constant, v_z is the velocity, and t is time. Since this is a second-order partial differential equation, the initial condition (C_A as a function of z) and two boundary conditions must be specified. Notice that, at steady state, this results in a second-order ordinary differential equation:

$$D_{Az} \frac{d^2 C_A}{dz^2} - v_z \frac{dC_A}{dz} - k C_A = 0 \quad (4)$$

which can be solved, given the boundary conditions.

III. PROCESS SIMULATION

Process simulation refers to the numerical solution of process models. At the process design stage, process simulation often refers to the numerical solution of an entire process flowsheet.

A. Algebraic Equations

The steady-state behavior of lumped parameter systems is characterized by a set of algebraic equations that have the form:

$$f(x) = 0 \quad (5)$$

obtained from Eq. (1) with a fixed p and u .

The most commonly used numerical techniques are related to Newton–Raphson iteration. The “guess” for iteration $k + 1$ is determined from the value at iteration k , using:

$$x(k+1) = x(k) - J(k)^{-1} f(x(k)) \quad (6)$$

where $f(x(k))$ is the vector of function evaluations at iteration k , and $J(k)$ is the Jacobian matrix:

$$J_{ij}(k) = \frac{\partial f_i}{\partial x_j}(k) \quad (7)$$

The ij element of the Jacobian represents the partial derivative of equation i with respect to variable j . If analytical derivatives are not available, elements of the Jacobian are obtained by perturbation of the state variable, requiring $n + 1$ function evaluations for an n -equation system of equations. Various quasi-Newton techniques provide approximations to the Jacobian and do not require as many function evaluations, thus reducing computational time.

In practice, the Jacobian matrix is not inverted; rather, a set of linear algebraic equations is solved for $x(k + 1)$

$$J(k)(x(k + 1) - x(k)) = -f(x(k)) \quad (8)$$

Some process models have more than one feasible solution. Most numerical methods have local convergence, so the solution obtained is dependent upon the initial guess for the solution before the first iteration. There is an ongoing effort to develop techniques that have global convergence or to find all solutions to multisolution problems.

Some chemical process systems may have a single steady state (single solution to a process model) under some design or operation conditions and multiple solutions under other design conditions. There are automatic techniques to vary a parameter of a system model to determine when these solutions branch from a single solution to multiple solutions. The FORTRAN code AUTO is perhaps the most widely used code for this.

A dynamic bifurcation occurs when the dynamic behavior of the solution to a system undergoes a qualitative change. For example, a subcritical Hopf bifurcation occurs when a dynamic system changes from a stable node to a limit cycle. Again, AUTO can be used to determine parameter changes that cause this bifurcation to occur.

B. Ordinary Differential Equations

Here we consider initial-value, ordinary differential equations which often arise when modeling time-dependent behavior of perfectly mixed systems. The general form is

$$\begin{aligned} \dot{x} &= \frac{dx}{dt} = f(x) \\ x_0 &= x(0) \end{aligned} \quad (9)$$

The explicit Euler integration technique involves specifying the integration step size, Δt :

$$x(k + 1) = x(k) + \Delta t f(x(k)) \quad (10)$$

This method often requires very small integration step sizes to obtain a desired level of accuracy. Runge–Kutta integration has a higher level of accuracy than Euler. It is also an explicit integration technique, since the state values at the next time step are only a function of the previous time step. Implicit methods have state variable values that are a function of both the beginning and end of the current

integration time. These methods often require several numerical iterations (usually a nonlinear algebraic equation is solved) for each integration step; Gear's method is such a procedure.

Most commonly used ordinary differential equation (ODE) solvers provide options of several different integration techniques. Most solvers also automatically vary the integration step size during the simulation to allow the best trade-off between accuracy and solution time, based on user-specified numerical tolerances. There is no single best integration technique—different methods work better for various problems.

So-called “stiff” differential equation models are particularly challenging to solve. Stiff models have dynamic behavior that encompasses a wide range of time scales. An example would be fast kinetics combined with long fluid-residence times in a chemical reactor. Gear's method is perhaps the most commonly used technique for solving these types of problems.

Differential algebraic equations commonly arise when physical property or kinetic expressions must be evaluated in dynamic problems. These systems have the following form:

$$M\dot{x} = M \frac{dx}{dt} = f(x)$$

where M is possibly singular. The most commonly used software code to solve these types of problems is DASSL.

C. Partial Differential Equations

A common method for solving partial differential equations (PDEs) is known as the “method of lines.” Here, finite difference approximations for spatial derivatives are used to convert a PDE model to a large set of ordinary differential equations, which are then solved using any of the ODE integration techniques discussed earlier.

Typically, the numerical solutions techniques used are very specific to the problem. Particularly challenging problems include “moving front” problems where concentration profiles, for example, may vary widely over a short distance but may not change much at other spatial locations. The spatial discretization must be small close to the front for accuracy and numerical stability, but must be larger at other locations to reduce computation time. Various adaptive grid techniques to change the spatial step sizes have been developed for these problems. One of the more common codes to solve fluid-flow-related problems is FLUENT.

In general, the numerical solution of PDEs is much more difficult to automate than the solution of initial-value ODEs. The best method to be used is very dependent on the problem being solved.

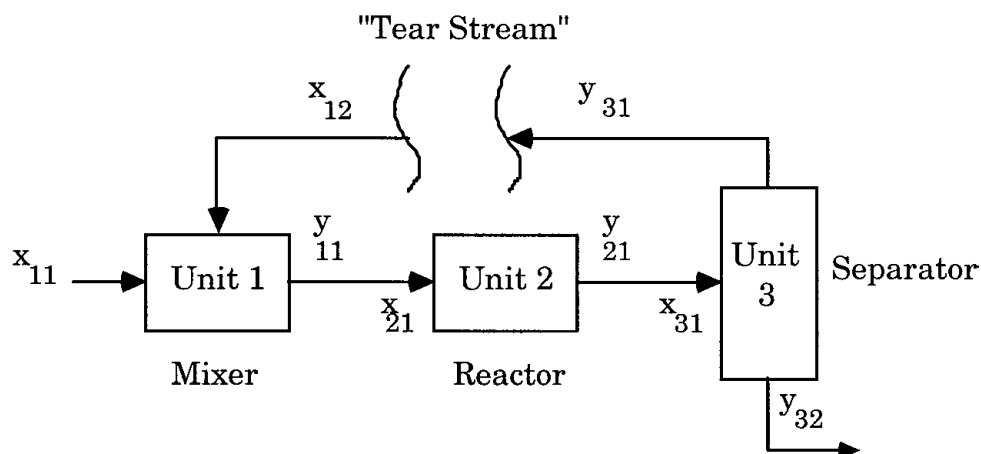


FIGURE 2 Sequential modular solution of a chemical process with recycle.

D. Flowsheet Simulations

Computers were not commonly used for simulation in the chemical process industry until the late 1950s, and few process engineers had access to the mainframe computers used at that time. Many oil and chemical companies and engineering firms began writing software to solve individual unit operations, such as distillation columns. Gradually, the stand-alone software codes were integrated so unit operation simulation modules could be solved sequentially to simulate process flowsheets. This method, known as the *sequential modular* method, is still widely used in commercial simulation packages to this day. An example of an early flowsheet simulator is FLOWTRAN from Monsanto (the former basic chemicals division of Monsanto is now known as Solutia).

By the late 1970s most companies decided it was not effective to continue development and support of their own in-house packages. Widely used commercial simulation packages included PROCESS from Simulation Sciences and DESIGN from CHEMSHARE. The ASPEN (Advanced System for Process Engineering) project at MIT was supported by the National Science Foundation to develop advanced computational software for chemical process simulation and design. The ASPEN package was further developed by ASPEN Technology, Inc., and the current release is known as ASPEN PLUS.

By the early 1990s, advances in user interfaces made the simulation packages much easier to use. HYSIM from Hyprotech was the first package to be developed specifically for a personal computer (PC) platform. All commonly used process simulators (including PRO/II from Simulation Sciences) are now available on a PC platform, with interfaces that allow flowsheet simulations to be formulated rapidly. Icons representing feed and product streams and unit operations are placed directly on a sim-

ulation flowsheet. Expert interfaces prompt the user for required information until the degrees of freedom have been completely specified.

The solution of a chemical process simulation problem using the sequential modular technique is represented in Fig. 2. Here, the modeling equations can be written such that the outlet stream from each unit is a function of the inlet streams to each unit:

$$y_{11} = g_{11}(x_{11}, x_{12})$$

$$y_{21} = g_{21}(x_{21}, p_2)$$

$$y_{31} = g_{31}(x_{31}, p_3)$$

$$y_{32} = g_{32}(x_{31}, p_3)$$

where x_{ij} = input j to unit i , y_{ij} = output j from unit i , and p_i = parameters for unit i .

In Fig. 2, the recycle stream has been selected as the "tear stream." A guess for x_{12} must first be made, then the equations for units 1, 2, and 3 are solved. The output of unit 3, y_{31} , is then compared with the original guess for x_{12} . The problem is solved when y_{31} has converged to x_{12} within the desired tolerance. The nonlinear algebraic equations to be solved can be written as $x_{12} = g(x_{12})$ or $f(x_{12}) = x_{12} - g(x_{12}) = 0$ and solved using the techniques discussed in Section III. Notice that the streams between units 1 and 2 or 2 and 3 could also have been chosen as the tear stream in Fig. 2.

An alternative to the sequential modular approach is to solve the equations modeling all of the units in a process flowsheet simultaneously; this is known as the *equation-based* approach. Advantages to the sequential modular approach include: (1) specialized numerical techniques tailored to each unit operation can be used, and (2) the numerical failure of one unit operation may still yield usable flowsheet information. Advantages to the equation-based

approach include: (1) recycle systems are more easily handled, and (2) flowsheet optimization is easier.

IV. OPTIMIZATION

A. Introduction

The petrochemical industry has undergone incredible changes during the past 25 years, due to increasingly strict environmental regulations, increases in raw material and energy costs, as well as intense competition from multinational competitors. Improvements in productivity have been impressive, driven in large part by improved operation strategies and process designs. Optimization is perhaps the most important tool utilized for these improvements.

What is meant by optimization? Optimization is the field of study associated with finding the best solution to mathematically defined problems. Fields that are impacted by optimization include physics, biology, engineering, economics, and mathematics:

- Engineering and semiconductor design
- Science and sequencing of the human genome
- Economics and comparing unemployment rate vs. inflation rate

Most of the optimization techniques in use today have been developed since the end of World War II. Considerable advances in computer architecture and optimization algorithms have enabled the complexity of problems that are solvable via optimization to steadily increase. Initial work in the field centered on studying linear optimization problems (linear programming, or LP), which is still used widely today in business planning. Increasingly, nonlinear optimization problems (nonlinear programming, or NLP) have become more and more important, particularly for steady-state processes.

Optimization can be performed on many different time scales and levels, from production planning over the next year to determining optimal setpoints for a chemical process unit operation every minute. Typical optimization levels in the petrochemical industry include management decisions, process design, and plant operations. In these cases, the solution to the optimization problem will be the one that maximizes some measure of profit. An example of optimization applied to process design is determination of the optimum thickness of insulation for a given steam pipe installation, as shown in Fig. 3.

The fixed costs increase as the insulation thickness is increased; however, the costs associated with heat losses decrease. The total cost goes through a minimum at the

optimum insulation thickness. Our discussion will focus on understanding the basic structure of optimization problems, explore some of the common solution techniques, and examine some of the recent directions/applications of optimization.

B. Structure of Optimization Problems

Optimization problems are by their nature mathematical in nature. The first and perhaps the most difficult step is to determine how to mathematically model the system to be optimized (for example, paint mixing, chemical reactor, national economy, environment). This model consists of an objective function, constraints, and decision variables. The objective function is often called the *merit* or *cost function*; this is the expression to be optimized that is the performance measure. For example, in Fig. 3 the objective function would be the total cost. The constraints are equations that describe the model of the process (for example, mass balances) or inequality relationships (insulation thickness >0 in the above example) among the variables. The decision variables constitute the independent variables that can be changed to optimize the system.

We can show the basic concepts and structure of optimization problems by examining a least squares problem. The problem in this case is to determine the two coefficients (α_0 and α_1) such that the error between the measured output and model predicted output is minimized:

$$\min_{\{\alpha_0, \alpha_1\}} f = \sum_{j=1}^p \varepsilon_j^2 \quad (11)$$

$$\varepsilon_j = Y_j - \hat{Y}_j \quad (12)$$

$$\hat{Y}_j = \alpha_0 + \alpha_1 x_j \quad (13)$$

In this example, the objective function to be optimized is given by Eq. (11), while the constraints that constitute the model are given in Eqs. (12) and (13). The number of measurements is p in Eq. (11) ε_j is the error between the measured and model predicted outputs, x is the independent variable, Y_j is the actual output, and $\hat{Y}_j$ is the model predicted output.

A number of steps are involved in the solution of optimization problems, including analyzing the system to be optimized so that all variables are characterized. Next, the objective function and constraints are specified in terms of these variables, noting the independent variables (degrees of freedom). The complexity of the problem may necessitate the use of more advanced optimization techniques or problem simplification. The solution should be checked and the result examined for sensitivity to changes in the model parameters.

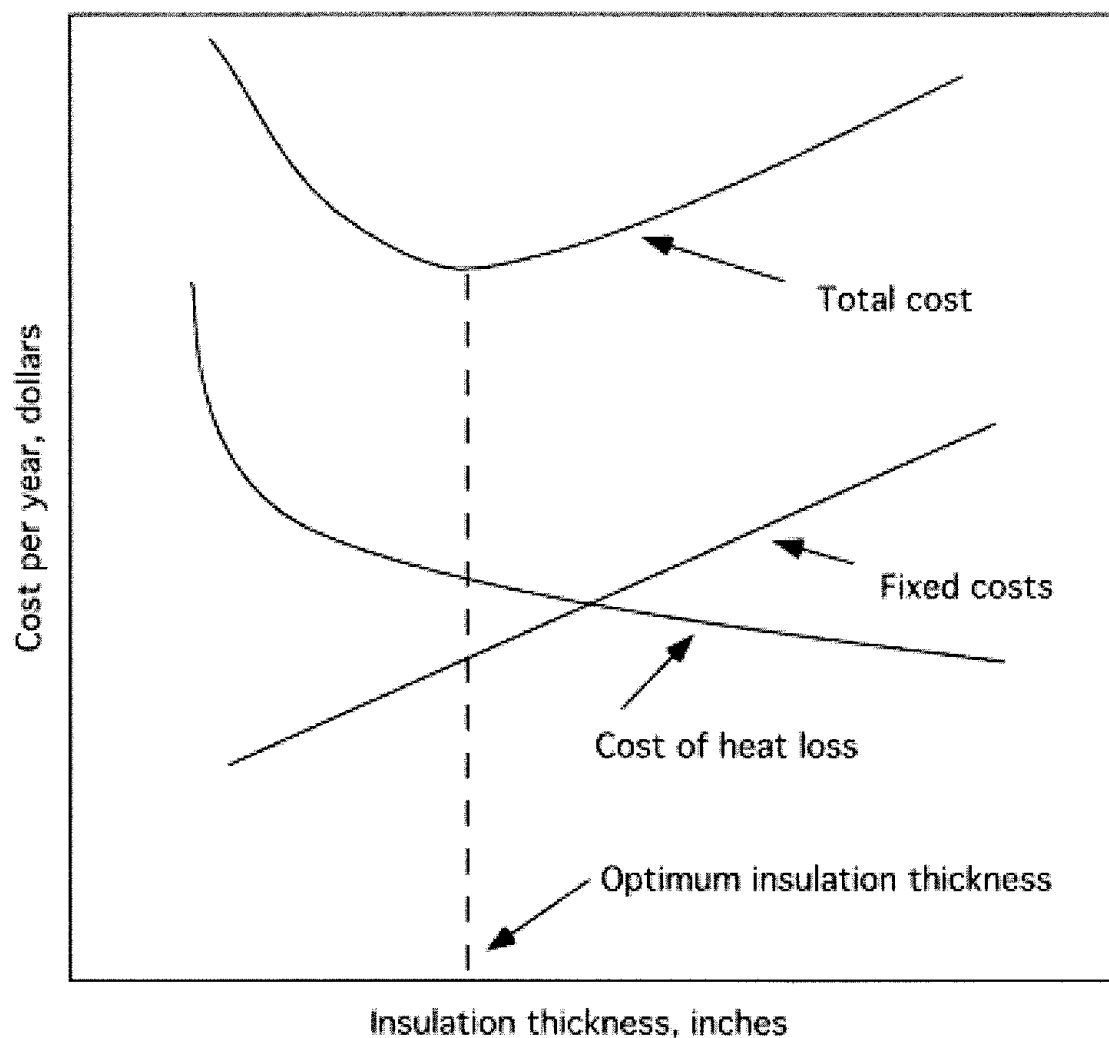


FIGURE 3 Determination of optimum insulation thickness.

C. Unconstrained Optimization

Unconstrained optimization deals with situations where the constraints can be eliminated from the problem by substitution directly into the objective function. Many optimization techniques rely on the solution of unconstrained subproblems. The concepts of convexity and concavity will be introduced in this subsection, as well as discussing unimodal versus multimodal functions, single-variable optimization techniques, and examining multi-variable techniques.

Multimodal functions are functions that have multiple minima/maxima over the independent variable range. Since most optimization techniques in use today search for local optima (minima or maxima), one can see that these techniques could easily fail to find the true optimum, also known as the global extrema. Multimodal functions are

characterized by their stationary points. Stationary points of a function are the values of the dependent variable (x) where the first derivative of the function (termed the gradient) is zero.

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} \quad (14)$$

Stationary points can be a (1) local maximum, (2) local minimum, or (3) saddle point. The existence of a stationary point is a necessary condition for an optimum.

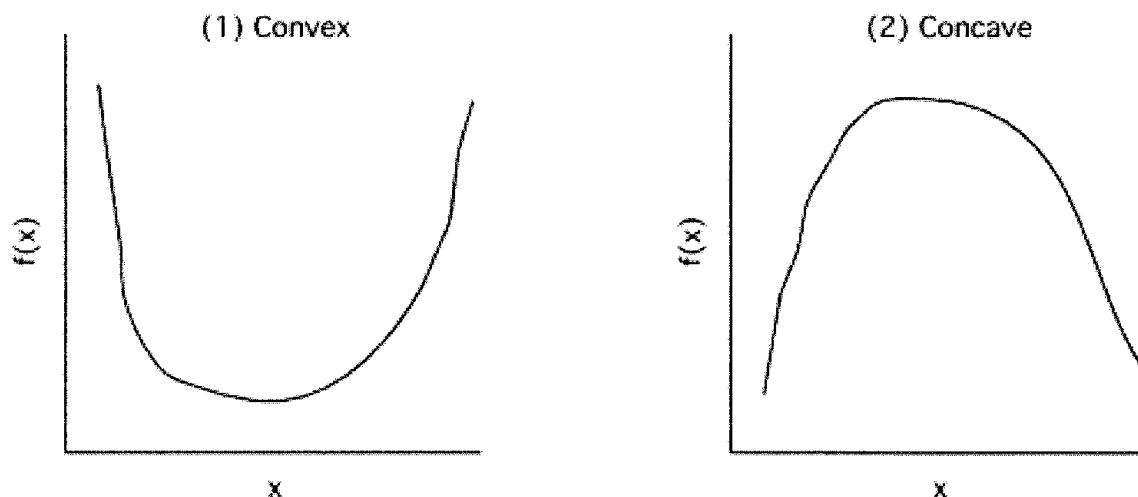


FIGURE 4 (1) Convex and (2) concave functions.

The concept of concave versus convex functions allow us to quantify the behavior of functions analytically (see Fig. 4).

A single-variable function is convex if the second derivative is strictly positive over the range of the dependent variable, as shown in Fig. 4(1). As shown in Fig. 4(2), a function is concave when the second derivative is negative over the dependent variable range. For a multivariable function, the matrix of second derivatives (termed the *Hessian*, $H(x)$) is used to check the convexity (or concavity) of the function:

$$H(x) = \nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \quad (15)$$

Extrema (minima or maxima) of a function can be examined by checking the eigenvalues of the Hessian at its stationary points. If all the eigenvalues of the Hessian are positive (negative) at a stationary point, then the function f is at a local minimum (maximum). Likewise, if all the eigenvalues of the Hessian are positive for all x , then $f(x)$ is said to be strictly convex, with a global minimum at the stationary point.

Efficient single-variable numerical techniques for optimization are important beyond their implementation for one-dimensional problems because they form the basis of most multivariable techniques. Three classes of tech-

niques for one-dimensional searches are often mentioned (Edgar et al., 2001):

1. Indirect
2. Region elimination
3. Interpolation

Indirect methods solve the necessary conditions for an optimum (looking at the “shape” of the function) directly via iteration. Region elimination techniques such as Fibonacci and Golden Section searches use function evaluations only to delete a portion of the independent variable range at each iteration. Interpolation techniques use polynomial fitting (quadratic or cubic oftentimes) to predict the location of the optimum.

There are two types of unconstrained multivariable optimization techniques: those requiring function derivatives and those that do not. An example of a technique that does not require function derivatives is the sequential simplex search. This technique is well suited to systems where no mathematical model currently exists because it uses process data directly.

Some of the more common unconstrained multivariable techniques include steepest descent, conjugate gradient, Newton’s method, and quasi-Newton methods. Steepest descent algorithms use first-order derivative information to maximize the rate of change of the objective function. Steepest descent algorithms have been shown to be sensitive to function scaling in practice and may exhibit poor convergence properties as a result. Conjugate gradient methods are essentially an improved version of steepest descent techniques that combine current information about the gradient vector with gradient information from past iterations.

Newton's method and quasi-Newton techniques make use of second-order derivative information. Newton's method is computationally expensive because it requires analytical first- and second-order derivative information, as well as matrix inversion. Quasi-Newton methods rely on approximate second-order derivative information (Hessian) or an approximate Hessian inverse. There are a number of variants of these techniques from various researchers; most quasi-Newton techniques attempt to find a Hessian matrix that is positive definite and well-conditioned at each iteration. Quasi-Newton methods are recognized as the most powerful unconstrained optimization methods currently available.

D. Constrained Optimization

Constraints in optimization problems often exist in such a fashion that they cannot be eliminated explicitly—for example, nonlinear algebraic constraints involving transcendental functions such as $\exp(x)$. The Lagrange multiplier method can be used to eliminate constraints explicitly in multivariable optimization problems. Lagrange multipliers are also useful for studying the parametric sensitivity of the solution subject to the constraints.

In general, an optimization problem involving constraints has the form:

$$\begin{aligned} \min f(x_1, \dots, x_n) \\ h(x_1, \dots, x_n) = 0 \end{aligned} \quad (16)$$

where $f(x)$ is a nonlinear function to be minimized, and $h(x)$ is a vector of nonlinear functions denoting the equality and inequality constraints. The necessary conditions for a local minimum of a general nonlinear function are given by the Kuhn–Tucker conditions for optimality. These conditions are often the basis for the design and termination criteria for optimization algorithms.

In addition to a wide variety of problem types, there are three common types of constrained optimization problems that are typically of interest: linear programs (LPs), quadratic programs (QPs), and nonlinear programs (NLPs).

Linear programming is one of the most common optimization techniques applied. LPs are commonly used on production scheduling and resourcing problems. A linear program is a class of optimization problems where the objective function and constraints are linear. The objective function and constraints of a linear program are convex; therefore, a local optimum is the global optimum. In addition, LPs demonstrate the characteristic wherein the optimum solutions of LPs lie on a constraint

or intersection of constraints. The most common solution technique employed is the Simplex method. In recent years, primal–dual interior-point linear programming algorithms have been introduced that have more favorable solution properties when compared to the Simplex method.

Quadratic programming involves the optimization of a quadratic function subject to linear constraints. Oftentimes, a quadratic program is solved as a subproblem while solving general nonlinear programming problems. There are several techniques for solving nonlinear programming problems, including the generalized reduced gradient (GRG) as well as successive quadratic programming (SQP) approaches.

The application of mixed-integer linear (MILP) and nonlinear (MINLP) programming approaches is rapidly increasing in popularity. These problems involve constraints that take on integer values. Improvements in processing power and algorithmic stability as well as the study of hybrid systems have led to the increasing use of these techniques. Global optimization techniques are also increasing in use, particularly for solving steady-state optimization problems where locating all solutions is practical computationally.

V. PROCESS DESIGN

Process design is a broad area. At one extreme it can include, for example, the specification of a replacement heat exchanger for an existing process. On the other hand, it can involve the design of all process-related equipment for an entirely new (grassroots) plant. A more common situation is a “retrofit” of an existing plant, where significant process equipment modifications are being made.

A. Process Synthesis

The conceptual development of a typical chemical process remains somewhat of an art. Usually, experience with similar process plants leads to an initial process flowsheet. Parameter optimization on that flowsheet can be used to determine the best economic design. Other flowsheets can be generated by adding/removing unit operations and process streams and changing the structure of the process flowsheet. A simplified depiction of this synthesis process is shown in Fig. 5.

Early work in process synthesis focused on the solution of specific problems, such as the best sequence of distillation columns to perform separation of components in feedstreams into product streams. Another early problem was the synthesis of heat-exchanger networks.

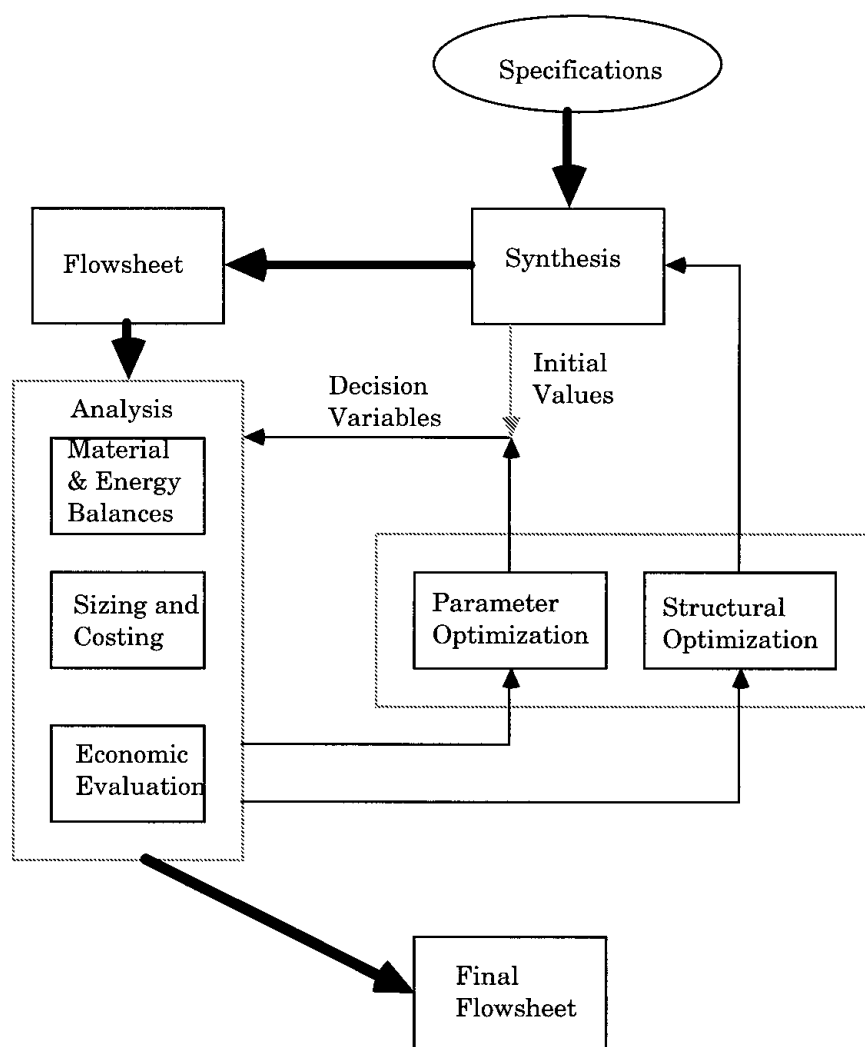


FIGURE 5 Decomposition of process design into related tasks. (Adapted from Westerberg et al., 1979.)

1. Sequence of Distillation Columns

Initial work in separations synthesis involved the best sequence of distillation columns to separate an n -component mixture. For example, consider the two possible distillation configurations to separate a three-component mixture (A, B, C ordered from light to heavy), shown in Fig. 6. The first, known as the "direct" sequence, yields a high-purity distillate product of the light component, followed by a column separating the two heavy components into high-purity streams. The alternative is to have a high-purity bottoms product stream from the first column, followed by a split of the light components into two high-purity streams. Clearly, the number of possible sequences increases dramatically with more feedstream components. The heuristics developed to rapidly screen for the best sequence work very well for ideal systems, where the

components are similar. More recent work has improved the selection of sequences for nonideal systems, such as those that form azeotropes. Finally, other sets of heuristics are available for the broader problem that includes extraction, crystallization, membranes, or chromatography-based separations processes.

2. Heat Exchanger Network Synthesis

Process plants normally have many streams that must be heated and many other streams that must be cooled. The heat-exchanger synthesis problem is to find a set of heat exchangers that minimize the total annual operating cost of the plant. In the past, dedicated utilities were used (steam for heating and cooling water for cooling), which resulted in systems that were easy to design and control but expensive to operate. A major driving force for increased

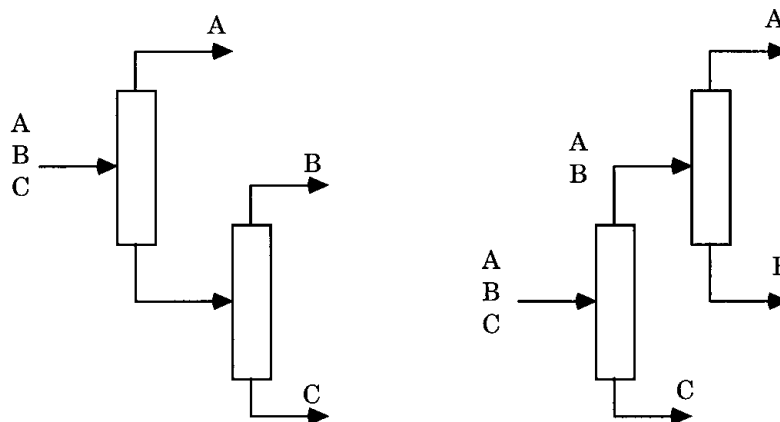


FIGURE 6 Two possible distillation sequences for a ternary feedstream.

energy integration was the energy crisis of the 1970s. So-called pinch technology was developed to determine the minimum consumption of utilities for a heat-exchanger network. One of the first commercially available codes was HEXTRAN from Simulation Sciences.

3. Task-Integrated Process Synthesis

An understanding of combined reaction and separations processes has led to process designs with significant capital and operation cost savings compared to more traditional process synthesis approaches. A prime example is the methyl acetate process patented by Eastman Kodak. A traditional methyl acetate process involves a reactor and nine separation vessels (including a mix of distillation and extraction operations). A revolutionary design resulted in a reactive/extractive distillation process with a single column. The resulting design yielded savings in both capital and operating costs of over 80%.

B. The Design Project

The focus of process engineers handling a chemical process design project is naturally on development of the process flowsheet. The process design team, using computer-aided simulations, hand calculations, and previous experience, specify the basic flows, heat duties, and separation stages and create the process flow diagram (PFD). Other engineers are often involved in the detailed equipment design. For example, the process team may specify an exchanger heat duty. A project engineer will perform the detailed exchanger equipment design, including size and number of tubes, materials of construction, etc. A good depiction of the total design project is shown in Fig. 7. Although the process design phase is less than 15% of the total project cost, it can result in a tremendous eco-

nomic impact. Different process designs can result in large differences in the project payout periods. It is particularly important for designs to have good flexibility (able to handle variations in process parameters, feed conditions, etc.) and operability (able to handle disturbances and be dynamically controllable). Dynamic simulation is being used to determine if a steady-state design is dynamically controllable or to test process safety in the event of equipment failure. Additionally, three-dimensional computer-aided design (CAD) is being used to enable the visualization of equipment layout (placement of valves, pipe racks, etc.) for ergonomic and safety reasons.

There has been active work in the development of processes that are safer and have less potentially damaging environmental impacts. These “environmentally benign” or “green” process designs typically include a life-cycle analysis to account for the long-term environmental (and economic) impact of a product or design.

VI. DESIGN FOR OPERABILITY AND FLEXIBILITY

A. Introduction

Increased process integration due to increases in energy and feedstock costs has made processes more difficult to operate. These operational difficulties have led to the need for integration of process design and process control. A number of techniques to address design for flexibility and operability have been proposed. These techniques include flexibility analysis (based on nonlinear programming and steady-state models), dynamic resiliency (based on linear, multivariable models), and steady-state and dynamic “back-off” analysis (where the actual operating point is chosen by “backing off” from the optimum point which lies at the intersection of constraints). Frequently,

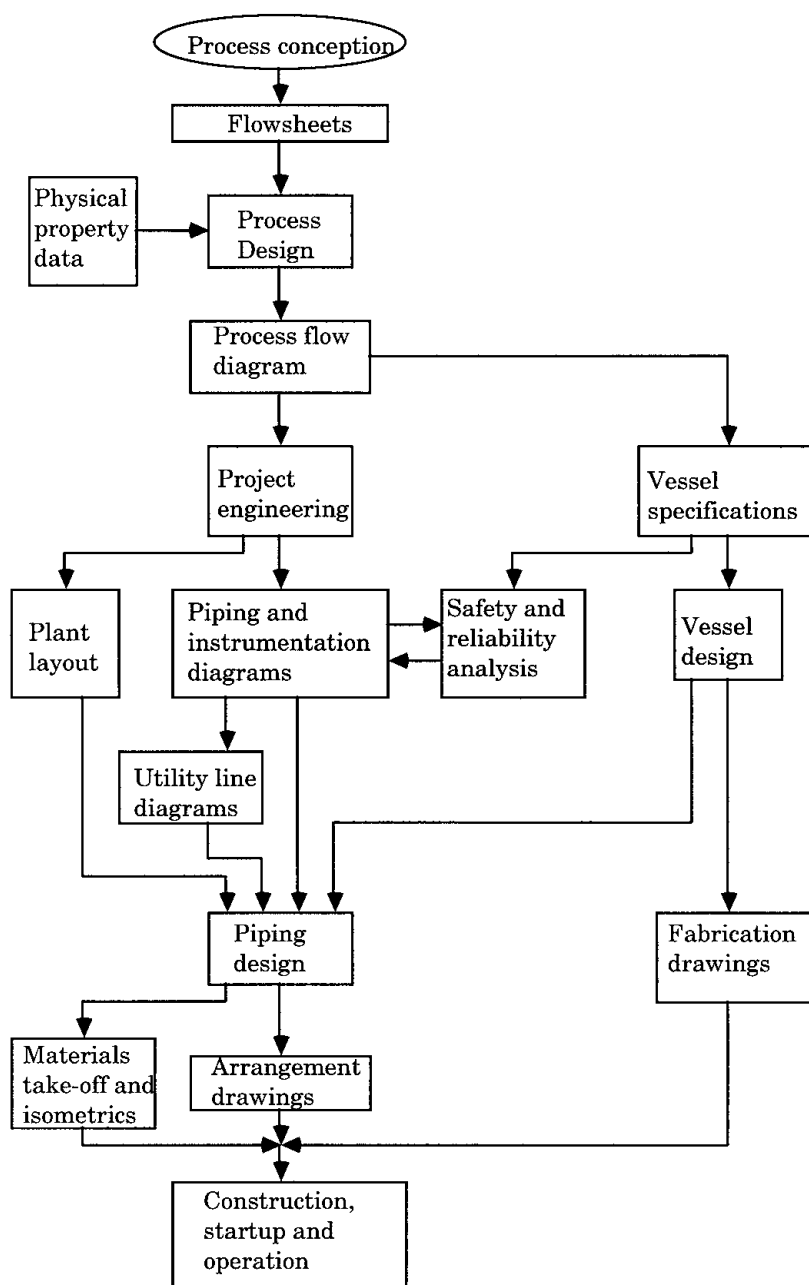


FIGURE 7 Total design project. (Adapted from Westerberg et al., 1979.)

operability and controllability improvements come at the cost of additional equipment and modifications to existing process designs. Capital costs are weighed against operating costs and other losses that result from nonflexible process designs.

A primary focus in process design is to produce designs that have good overall operability characteristics, as determined by safety and environmental considerations, low product variability, low operating costs, and ease of startup/shutdown. The process design and control system

design is iterative in nature, such that considerations of operability and dynamic controllability (through numerical simulation studies) may cause the process designer to modify the design. This in turn leads to processes that are capable of lower variability and hence improved product quality as measured by process capability indices. Process variability is minimized/distributed among all the unit operations (not just in the “end of the process”), and the resulting design is robust to deficiencies that may occur (heat-exchanger fouling, loss of equipment, etc.).

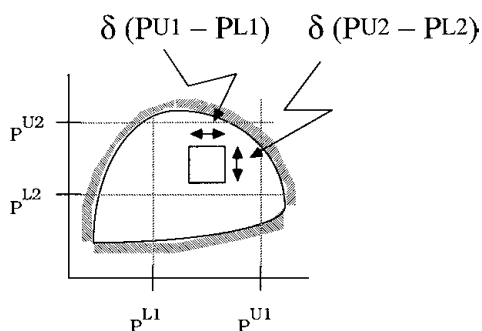


FIGURE 8 Schematic of the flexibility index.

Additionally, the increased use of model predictive control techniques allows for more degrees of freedom (associated with having more manipulated variables) which in turn tends to improve controllability and therefore decrease process variability. Principal component analysis (PCA) and projection to latent structures (PLS) are tools that can be used to monitor process performance and behavior.

B. Design for Flexibility

A process design is flexible if it can tolerate uncertainties in parameters and can handle disturbances. A flexibility index is a measure of the amount of uncertainty that can be tolerated with the desired process operation remaining feasible. A schematic of the flexibility index, δ , is shown in Fig. 8. Here, the two degrees of freedom represent uncertain parameters or disturbance variables, which have assumed upper and lower bounds. The feasible operating region lies within the cross-hatched area. The flexibility index is the fraction of the parameter range that still results

in feasible operation. This computation is reasonably simple for a convex problem, as illustrated, but becomes much more complicated for nonconvex problems and those with a large number of degrees of freedom.

C. Design for Operability

Process design modifications usually have a bigger impact on operability (dynamic resilience). Dynamic resilience depends on controller structure, choice of measurements, and manipulated variables. Multivariable frequency-response techniques have been used to determine resilience properties. A primary result is that closed-loop control quality is limited by system invertability (nonminimum phase elements). Additionally, it has been shown that steady-state optimal designs are not necessarily optimal in dynamic operation.

Another consideration is the dynamic controllability of a process design. If there are no uncertainties or disturbances, then the optimum economic design normally occurs at a constraint. An actual process cannot be operated at a constraint, because any disturbance may force the system to violate the constraint (a product purity limit, for example). In this case, the desired steady-state operating point must be “backed-off” from the economic optimum so that the control strategy can tolerate disturbances. The basic idea is shown in Fig. 9.

VII. PROCESS OPERATION

The actual operation of a chemical process involves decisions and actions that occur at a number of levels and

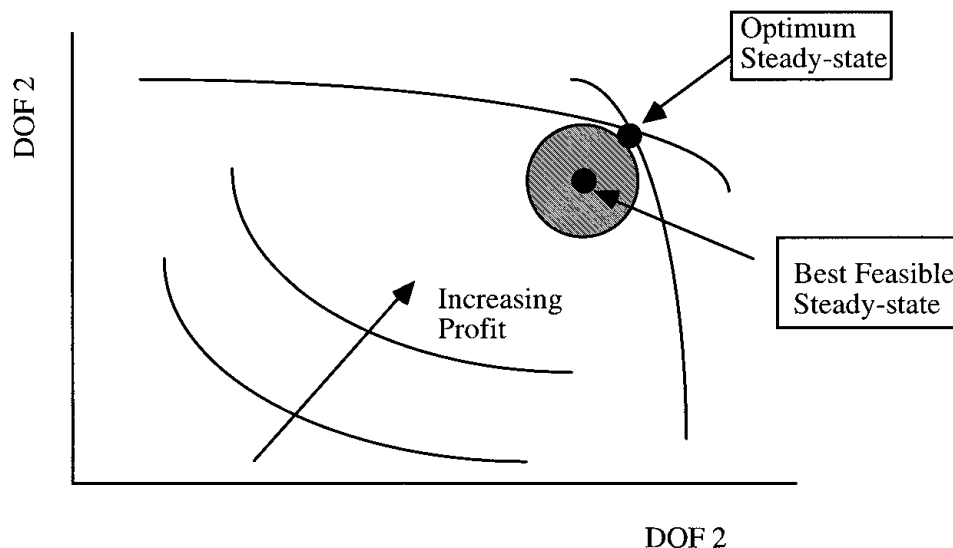


FIGURE 9 Illustration of best feasible steady-state operating point. (Adapted from Morari and Perkins, 1995).

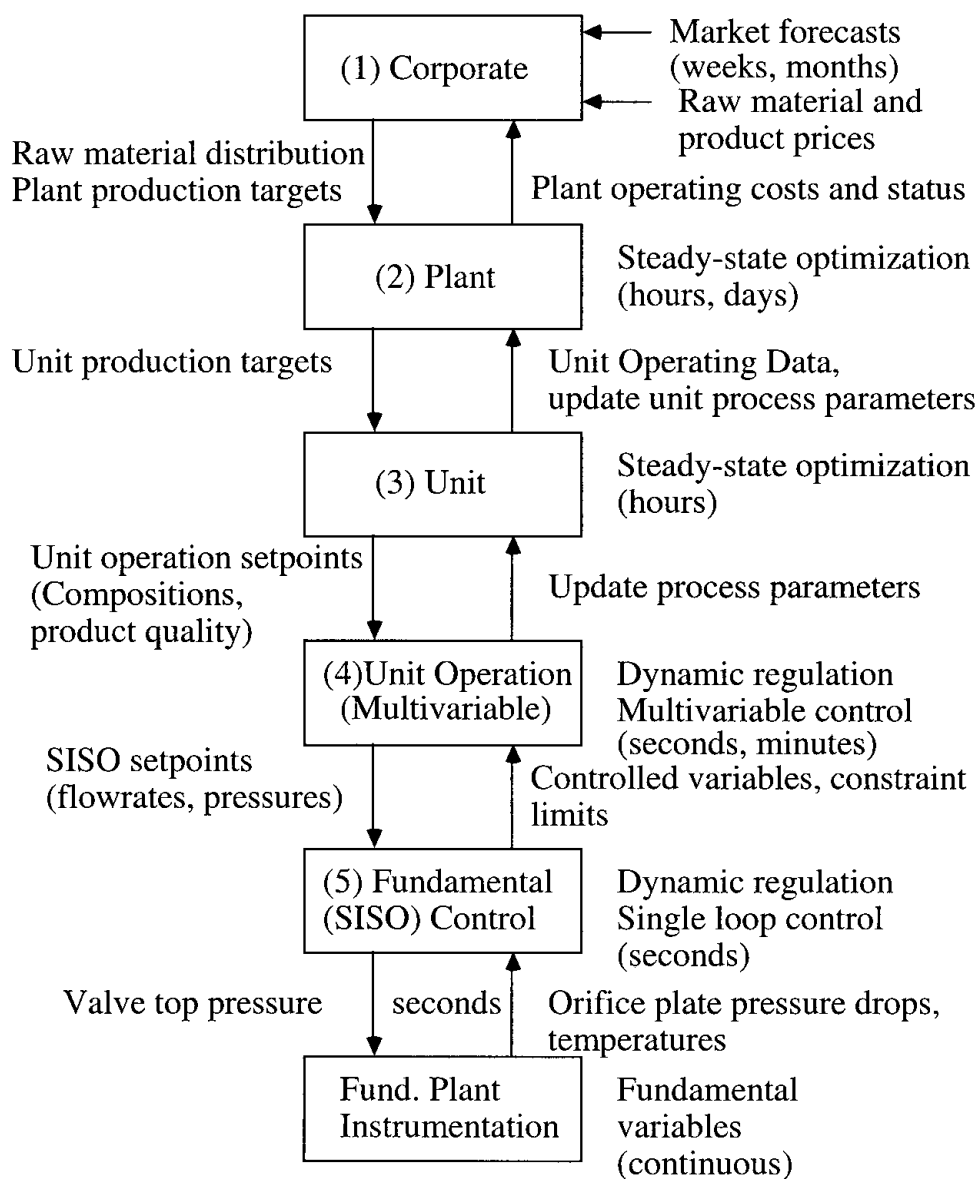


FIGURE 10 Operating levels.

time scales, as shown in Fig. 10. At the *corporate* level, allocation decisions are made. Market demand projections, raw material availability, and past operating costs are used to set long-term plans for corporate-wide optimization. These decisions are typically made on an infrequent basis—weekly, monthly, or quarterly. Often these decisions are made using linear programming techniques. This information is passed to the *plant* level.

At the next level are plant decisions about allocation among various plant operating units. These may be made daily or even more frequently depending on the plant automation and control system. For example, steam may be generated by a number of different sources, with different

costs depending on relative energy and product values. In an oil refinery, different units make similar components for blending into gasoline (which is produced in many different grades), and these relative amounts may change depending on energy and product values, operating problems with certain units, etc.

The *operating unit* (level three) takes information available on a daily basis and makes hour-by-hour changes to meet those goals, on the average, over the course of the day. The *unit operation* level involves individual equipment, and desired setpoints may be changed on an hourly (or more frequent) basis. Regulation on this level occurs on the order of minutes.

The *fundamental control* level consists primarily of process flow or pressure controllers. The lowest level generally involves the manipulation of a control valve or the reading of a sensor, where the time frame is of the order of one second or less.

Notice that there is a feedback of information from the lower levels back up to the higher levels. The basic idea of feedback control occurs at the unit operation and fundamental control levels. Here, control algorithms are used to adjust manipulated inputs (controller outputs or control variables) to maintain process outputs (process variables) at desired values (known as setpoints). The basic principles and techniques of process control are pre-

sented in the article by Edgar and Hahn (Process Control Systems).

A. Petroleum Refining Example

An example of the overall operating problem is illustrated in Fig. 11 for a typical petroleum refining company. At the highest level, corporate management decides where to purchase crude and how to distribute the crude oil to the various refineries in the corporation. At the next level (level 2), Refinery A takes the current and future crude delivery projections and gasoline production projections and determines the operating conditions for each process unit

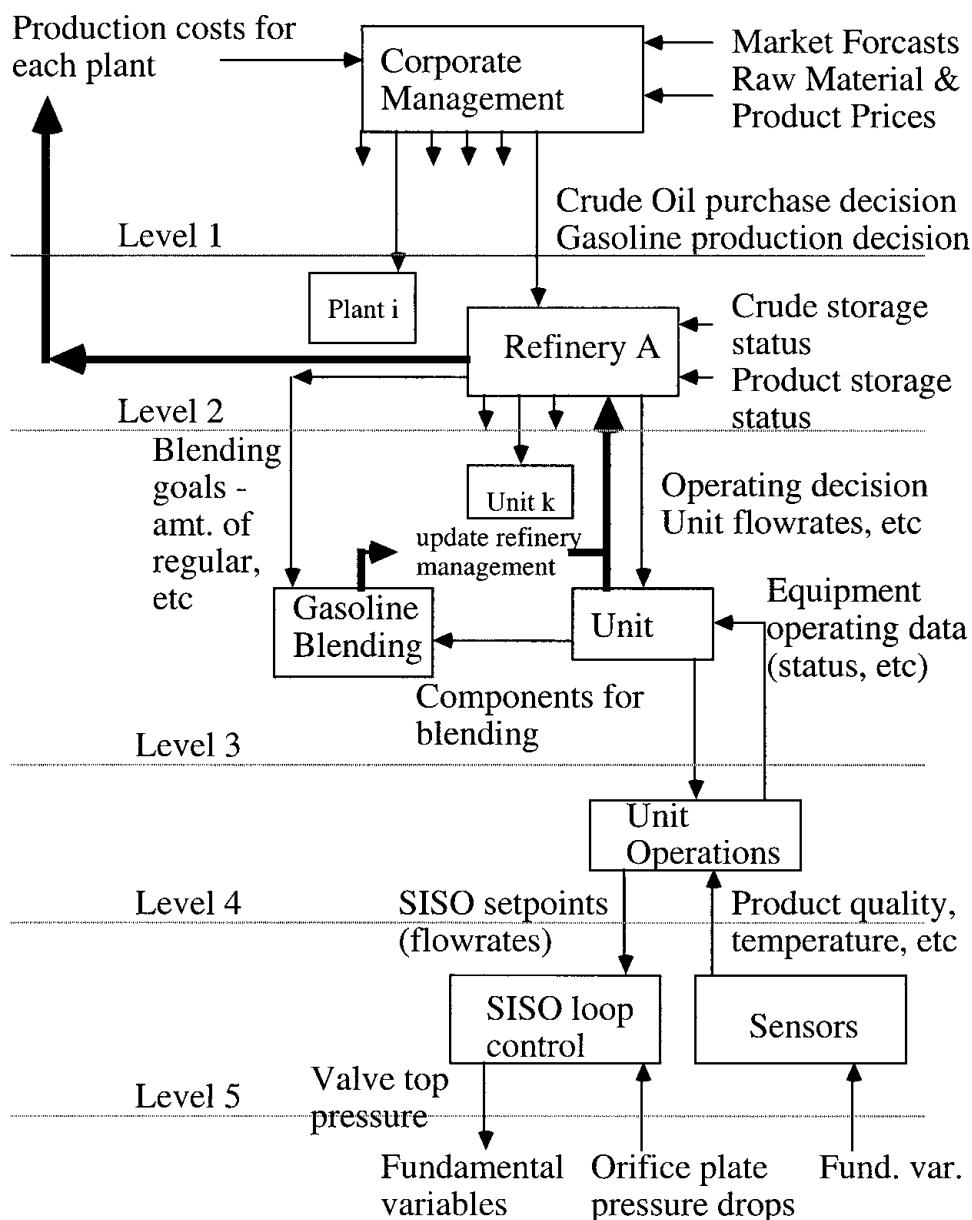


FIGURE 11 Petroleum refining optimization.

(level 3; e.g., the catalytic reforming unit) in the refinery. Setpoints for the unit operations (level 4; e.g., distillation) are determined at this point. The unit operations level determines the process flowrates, such as the distillate or reflux flowrates (level 5). These controllers then determine, for example, the pressure to the control valves to regulate various flowrates.

B. Real-Time Optimization

It should be noted that the optimization problems solved for levels 2 and 3 begin to merge as the plantwide optimization begins to set targets for the unit operations in many process units. This large-scale, frequent optimization of operating conditions is known as real-time optimization (RTO). RTOs are run approximately every 30 minutes to 1 hour, with the resulting optimal setpoints downloaded to model predictive controllers (MPC).

C. Data Reconciliation

Another important optimization problem that is solved frequently is data reconciliation. All measurements have some degree of uncertainty, and the measurements need to be reconciled so that the entire set of measurements is consistent with plant material and energy balances. This is particularly useful for monitoring inventories and for improving model predictions used in MPC.

D. Vertical Integration of Software and Consulting Firms

Until relatively recently the wide range of process systems-related activities was performed by a wide variety of independent design and consulting services firms. Simulation software was provided by a number of software companies, and there were many different software packages for particular simulations. Process designs were often performed by companies working in specific process areas (catalytic reforming, etc.). Project engineering firms provided more detailed and integrated plant design layout, developed the overall piping and instrumentation diagrams, and contracted the equipment fabrication, while construction companies supervised the construction project. Control and automation consulting firms provided software for control and assisted with process startup.

In recent years, however, there has been a vertical integration of process systems-related engineering services. Computer software firms have purchased or merged with control and automation companies and process control consulting firms. It is now becoming feasible to use the same software, or have relatively transparent links that allow a smooth transition between simulation models for design, data reconciliation, and parameter estimation, on-line optimization, and control.

SEE ALSO THE FOLLOWING ARTICLES

CERAMICS, CHEMICAL PROCESSING OF • HEAT EXCHANGERS • MINERAL PROCESSING • PETROLEUM REFINING • PHARMACEUTICALS • PLASTICS ENGINEERING • POLYMER PROCESSING • PROCESS CONTROL SYSTEMS • PULP AND PAPER • TEXTILE ENGINEERING

BIBLIOGRAPHY

- Bequette, B. W. (1998). "Process Dynamics: Modeling, Analysis and Simulation," Prentice Hall, Upper Saddle River, NJ.
- Biegler, L. T. (1989). "Chemical process simulation," *Chemical Engineering Progress* October, 50–61.
- Biegler, L. T., Grossmann, I. E., and Westerberg, A. W. (1997). "Systematic Methods of Chemical Process Design," Prentice Hall, Upper Saddle River, NJ.
- Douglas, J. M., and Stephanopolous, G. (1995). "Hierarchical approaches in conceptual process design: framework and computer-aided implementation," In "Foundations of Computer-Aided Process Design" (L. T. Biegler and M. F. Doherty, eds.), pp. 183–197, AIChE Symposium Series, **91**(304), New York.
- Edgar, T. F. (2000). "Process information: achieving a unified view," *Chemical Engineering Progress* **96**(1), 51–57.
- Edgar, T. F., Himmelblau, D. M., and Lasdon, L. S. (2001). "Optimization of Chemical Processes," McGraw-Hill, New York.
- Fletcher, R. (2000). "Practical Methods of Optimization," 2nd ed., Wiley, Chichester.
- Malone, M. F., Trainham, J. A., and Carnahan, B., eds. (2000). "Foundations of Computer-Aided Process Design," AIChE Symposium Series, **96**(323), New York.
- Morari, M., and Perkins, J. (1995). "Design for operations," In "Foundations of Computer-Aided Process Design" (L. T. Biegler and M. F. Doherty, eds.), pp. 105–114, AIChE Symposium Series, **91**(304), New York.
- Seider, W. D., Brengel, D. D., and Widagdo, S. (1991). "Nonlinear analysis in process design," *American Institute of Chemical Engineers Journal* **37**(1), 1–38.
- Seider, W. D., Seader, J. D., and Lewin, D. R. (1999). "Process Design Principles," Wiley, New York.
- Siirola, J. J. (1995). "An industrial perspective on process synthesis," In "Foundations of Computer-Aided Process Design" (L. T. Biegler and M. F. Doherty, eds.), pp. 222–233, AIChE Symposium Series, **91**(304), New York.
- Westerberg, A. W., Hutchison, H. P., Motard, R. L., and Winter, P. (1979). "Process Flowsheeting," Cambridge University Press, Cambridge, U.K.
- Wright, S. (1997). "Primal–Dual Interior-Point Methods," Society for Industrial and Applied Mathematics Philadelphia, PA.

The following are widely used numerical packages:

- ASPEN PLUS—process simulation
- AUTO—bifurcation analysis
- FLUENT—fluid-flow simulation
- HEXTRAN—heat-exchanger network synthesis
- HYSYS—process simulation
- PRO/II—process simulation



Coherent Control of Chemical Reactions

Robert J. Gordon

University of Illinois at Chicago

Yuichi Fujimura

Tohoku University

- I. Overview
- II. Mode-Selective Chemistry
- III. Coherent Phase Control
- IV. Wave Packet Control
- V. Control of External Degrees of Freedom
- VI. Concluding Remarks

GLOSSARY

Coherent control Control of the motion of a microscopic object by using the coherent properties of an electromagnetic field. *Coherent phase control* uses a pair of lasers with long pulse durations and a well-defined relative phase to excite the target by two independent paths. *Wave packet control* uses tailored ultrashort pulses to prepare a wave packet at a desired position at a given time.

Coherent population transfer Transfer of population from one quantum mechanical level to another using coherent radiation. The radiation may be provided by either continuous or pulsed lasers. Using the method of adiabatic passage (see *STIRAP*), 100% population transfer has been achieved.

Genetic algorithm A learning algorithm used to maximize the adaptability of a system to its environment. The method, based on the genetic processes of re-

production, crossover, and mutation, has been used to optimize the amplitudes and phases (the “genes”) of the frequency components of a laser pulse in order to generate a wave packet with desired chemical properties.

Mode-selective chemistry The use of laser beams to control the outcome of a chemical reaction by exciting specific energy states of the reactants.

Optimal control theory A method for determining the optimum laser field used to maximize a desired product of a chemical reaction. The optimum field is derived by maximizing the objective function, which is the sum of the expectation value of the target operator at a given time and the cost penalty function for the laser field, under the constraint that quantum states of the reactants satisfy the Schrödinger equation.

Pendular state Superpositions of field-free rotational eigenstates in which the molecular axis librates about the field direction. Pendular states are eigenstates of the rotational Hamiltonian plus the dipole potential.

STIRAP Transfer of population by means of Stimulated Raman Adiabatic Passage, using a pump and Stokes laser. Population in a three-level system is completely transferred without populating the intermediate state if the Stokes laser precedes the pump laser in a “counter-intuitive” order.

Wave packet A localized wave function, consisting of a non-stationary superposition of eigenfunctions of the time-independent Schrödinger equation.

COHERENT CONTROL refers to a process in which the coherent properties of an electromagnetic field are used to alter the motion of a microscopic object such as an electron, atom, or molecule. The controlled process may be categorized according to the degree of freedom that is manipulated. For example, a laser beam with carefully tailored properties might be used to control the motion of electrons within an atom or molecule, thereby populating specific eigenstates, or to create electronic wave packets with interesting spatial and temporal properties. Another possibility is the use of coherent light to control the stretching and bending modes of a molecule, thereby altering its chemical reactivity. These are both examples of the control of *internal* degrees of freedom. Alternatively, a coherent light source might be used to orient a molecule in space so that a particular bond is pointing in a chosen direction. Another possibility is to use a focused laser beam to control the translational motion of an atomic or molecular beam, perhaps focusing the particles to a small volume or steering them in a new direction. In a condensed phase, a laser might be used to alter the direction of an electric or ion current. These are illustrations of control of *external* degrees of freedom. In all of these examples, the coherence of a light wave is transferred to a material target so as to alter the dynamical properties of the target in a controlled way.

I. OVERVIEW

In this article we approach the topic of coherent control from the perspective of a chemist who wishes to maximize the yield of a particular product of a chemical reaction. The traditional approach to this problem is to utilize the principles of thermodynamics and kinetics to shift the equilibrium and increase the speed of a reaction, perhaps using a catalyst to increase the yield. Powerful as these methods are, however, they have inherent limitations. They are not useful, for example, if one wishes to produce molecules in a single quantum state or aligned along some spatial axis. Even for bulk samples averaged over many quantum states, conventional methods may be ineffective in maximizing the yield of a minor side product.

A number of strategies have been developed over the past few decades to overcome the limitations of bulk kinetics. One method, known as *mode-selective chemistry*, exploits the idea that a molecule may have an eigenstate that strongly overlaps a desired reaction coordinate. Depositing energy into that degree of freedom may selectively enhance the reaction of interest. In the following section we give a number of examples using localized nuclear or electronic motion to enhance a particular process. Although this approach does not depend intrinsically on the coherent properties of light and therefore lacks one of the characteristic features of coherent control, it has played an historic role in the development of control techniques.

An inherent limitation of mode-selective methods is that Nature does not always provide a local mode that coincides with the channel of interest. One way to circumvent the natural reactive propensities of a molecule is to exploit the coherence properties of the quantum mechanical wave function that describes the motion of the particle. These properties may be imparted to a reacting molecule by building them first into a light source and then transferring them to the molecular wave function by means of a suitable excitation process.

Two qualitatively different (though fundamentally related) strategies for harnessing the coherence of light were developed in the mid-1980s. The first, proposed and developed by Paul Brumer and Moshe Shapiro, is a molecular analogue of Young’s two-slit experiment, in which two coherent excitation paths promote the system to a common final state. Variation of the relative phase of the two paths produces a modulation of the excitation cross section. This method does not rely on the temporal properties of the light source and may in principle use a continuous laser. We refer to this approach as *coherent phase control* and describe it in detail in Section III.

The second approach, proposed by David Tannor and Stuart Rice, and further developed by them and others including Ronnie Kosloff and Herschel Rabitz, uses very short pulses of light to prepare a wave packet that evolves in time after the end of the pulse. After a suitable delay, an interrogating pulse projects out the product of interest. *Wave packet control* may be thought of as a generalization of mode-selective chemistry in which a short (and therefore broadband) pulse of light produces a localized non-stationary state that evolves in a predetermined fashion. Wave packet methods have been used with considerable success to control electronic, vibrational, and rotational motion of a variety of simple systems. One of the very powerful properties of this approach is that it is possible to use automated learning algorithms to tailor the laser pulses to create wave packets with desired properties. Details of wave packet control are given in Section IV.

Coherent radiation may also be used to control the *external degrees of freedom* of a molecule. For example, it is possible to create a quivering “pendular state” in which a molecule having an anisotropic polarizability is aligned along the electric field vector of a laser beam. It is also possible to use a focused laser beam to deflect a beam of molecules, perhaps focusing them to a point or steering them towards a target. Control of external degrees of freedom is discussed in Section V. We conclude this article with a brief discussion of future directions that the field is likely to take.

II. MODE-SELECTIVE CHEMISTRY

The central concept of mode-selective chemistry is illustrated in Fig. 1, which depicts the ground and excited state potential energy surfaces of a hypothetical triatomic molecule, ABC. One might wish, for example, to break selectively the bond between atoms A and B to yield products A+BC. Alternatively, one might wish to activate that bond so that in a subsequent collision with atom D the products AD+BC are formed. To achieve either goal it is necessary to cause bond AB to vibrate, thereby inducing motion along the desired reaction coordinate.

Direct excitation to the continuum usually (but not always, *vide infra*) results in rupture of the weakest bond. In order for the experimenter to have control over which bond is broken, it is helpful first to excite motion along the bond of interest. This process, known as vibrationally mediated photodissociation, preselects the desired degree of freedom before the reaction takes place. This method is illustrated in Fig. 1, where a low energy photon excites

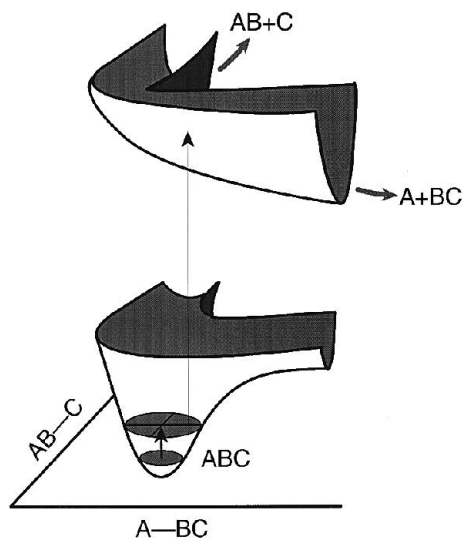


FIGURE 1 Illustration of mode-selective control of the dissociation of a triatomic molecule. (Provided by the courtesy of Fleming Crim.)

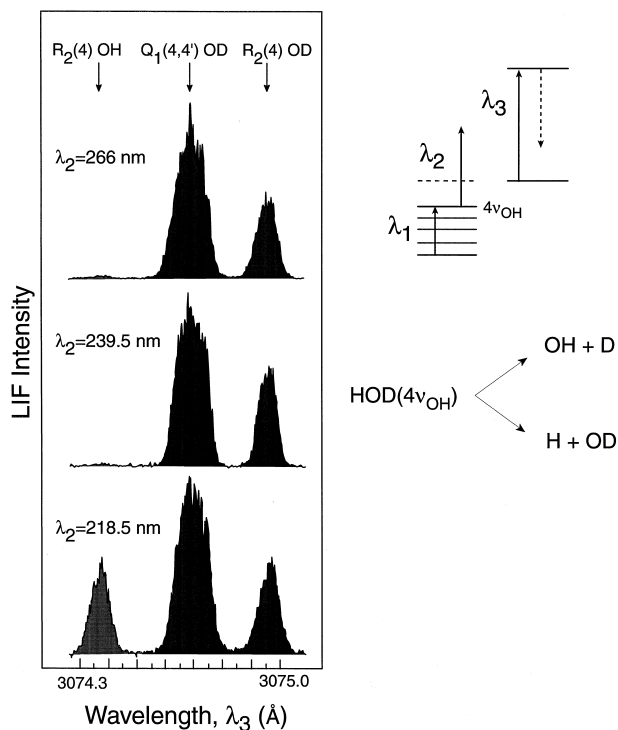


FIGURE 2 Vibrationally mediated photodissociation of water. (Provided by the courtesy of Fleming Crim.)

a bound state of the molecule causing the A–B bond to stretch. A high energy photon then promotes the molecule to the A+BC product valley of an excited potential energy surface.

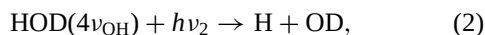
The possibility of such preselection depends on the local mode character of the molecule. Typical narrow-band light sources excite a stationary eigenstate of the Hamiltonian. Such an eigenstate may be written as a linear combination of zero-order states, ϕ_i , which correspond to localized motion such as stretches of individual bonds or simple bending motion. Our goal is to excite one of those zero-order states, such as the A–B stretch. In a favorable case, that zero-order state will carry most of the oscillator strength and will also be a major component of the excited eigenstate. Designating the wave function of the zero-order “bright” state by ϕ_s and all the other zero-order states as “dark,” we may write the excited state wave function as

$$\psi = c_s \phi_s + \sum_{i \neq s} c_i \phi_i. \quad (1)$$

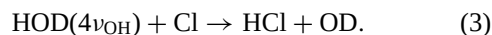
The bright-state character corresponding to localized vibration of the AB bond equals $|c_s|^2$. For anaharmonic molecules it is not uncommon to find eigenstates with large local mode character, i.e., with $|c_s|^2 \cong 1$.

Figure 2 illustrates the vibrationally mediated bond-specific photodissociation of isotopically labeled water,

HOD. A 722.5 nm laser pulse (λ_1) excites the third overtone stretch of OH. After a short delay, a pulse of ultraviolet radiation of frequency ν_2 (wavelength λ_2) dissociates the molecule, and a third pulse with a wavelength near 308 nm (λ_3) probes the OH or OD fragments by laser-induced fluorescence. It is observed that with a dissociation wavelength of 266 or 239.5 nm, the products are almost exclusively H + OD,



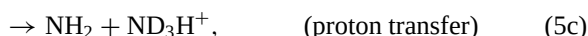
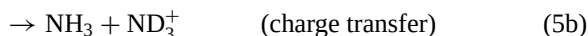
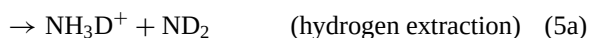
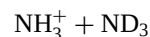
as seen in the $Q_1(4,4')$ and $R_2(4)$ fluorescence lines of OD, whereas with 218.5 nm equal amounts of OH and OD are formed. Because a stationary state of the molecule is excited by the first laser, the bond remains energized indefinitely until it collides with another particle. The excited molecule can then react, breaking preferentially the activated bond. For example, collision of $\text{HOD}(4\nu_{\text{OH}})$ with a chlorine atom produces primarily HCl rather than DCl:



The same principles have been applied to molecules with four atoms. For example, Fleming Crim and coworkers showed that excitation of the NH stretch of isocyanic acid enhances its reaction with Cl atoms,



whereas excitation of the bending mode inhibits the reaction. For the reactions of ammonia ions with neutral ammonia molecules,



Richard Zare and coworkers found that excitation of the umbrella mode of NH_3^+ selectively enhances the proton transfer reaction; however, in this case the projection of the nuclear motion onto the reaction coordinate is not as obvious.

Vibrational mode selectivity can also be used to promote electronic processes. Vibrational autoionization is a process whereby a bound electron acquires sufficient energy to escape by extracting one quantum of vibrational energy from the ionic core of the molecule. For such an energy transfer to occur, the electron must first collide with the core. Scattering of the electron with the core can be promoted if the amplitude of the nuclear motion overlaps the electronic charge density. An example of this process studied by Steven Pratt is vibrational autoionization of the 3d Rydberg electrons of ammonia, which is enhanced

by the umbrella vibration of the molecule. The d_{z^2} , d_{xz} , and d_{yz} orbitals have lobes perpendicular to the plane of the $\bar{C}'$ Rydberg state of NH_3 . Excitation of the out-of-plane umbrella mode of the molecule promotes vibrational autoionization of electrons in these orbitals but has little effect on the poorly overlapping $d_{x^2-y^2}$ and d_{xy} electrons.

It is also possible to use localized *electronic* excitation to promote reactions selectively. An example studied by Laurie Butler and coworkers is the ultraviolet photodissociation of CH_2IBr . This molecule has absorption maxima at 270, 215, and 190 nm, corresponding to localized excitation of a nonbonding iodine electron to an antibonding orbital localized on the C–I bond, ($n_1 \rightarrow \sigma_{\text{C-I}}^*$), promotion of a nonbonding bromine electron ($n_{\text{Br}} \rightarrow \sigma_{\text{C-Br}}^*$), and a Rydberg transition, respectively. Photodissociation at 248.5 nm, at the edge of the $n_1 \rightarrow \sigma_{\text{C-I}}^*$ transition, yields 60% I atoms and 40% Br. At 210 nm only Br atoms are formed, even though the C–I bond is the weakest bond in the molecule. In addition, some concerted IBr elimination occurs. At 193 nm all three products are formed.

Another example of electronic control studied by Butler is the photodissociation of methyl mercaptan, CH_3SH . Although the $\text{CH}_3\text{--SH}$ bond is the weakest bond in the molecule, $\text{CH}_3\text{S} + \text{H}$ are the primary photodissociation products at 193 nm. Bond selectivity in this case occurs even though the initially excited state is not repulsive along the reaction coordinate. Here selectivity results from non-adiabatic coupling of the initially excited metastable $2^1A''$ Rydberg state to the dissociative $n_1 \rightarrow \sigma_{\text{S-H}}^*$ state. Another case where nonadiabatic coupling results in bond-selective chemistry is the photodissociation of bromoacetyl chloride, BrCH_2COCl . It was found for this molecule that at 248 nm the C–Cl bond is preferentially broken, even though the barrier for C–Br scission is lower than that for C–Cl scission. The reason for bond selectivity in this case is that the splitting between the adiabatic potential energy surfaces is much smaller for C–Br scission, so that nonadiabatic crossing and recrossing without reaction is much faster in this channel as compared with adiabatic motion along the C–Cl reaction coordinate. These examples illustrate that, although bond-selective photoexcitation is a general phenomenon, its mechanism depends strongly on the details of the potential energy surfaces. Studies of mode-selective reactions are, therefore, a valuable source of information about the structure of potential energy surfaces and their interactions.

III. COHERENT PHASE CONTROL

The underlying principle of coherent phase control is that the probability of an event occurring is given by the square of the sum of the quantum mechanical amplitudes

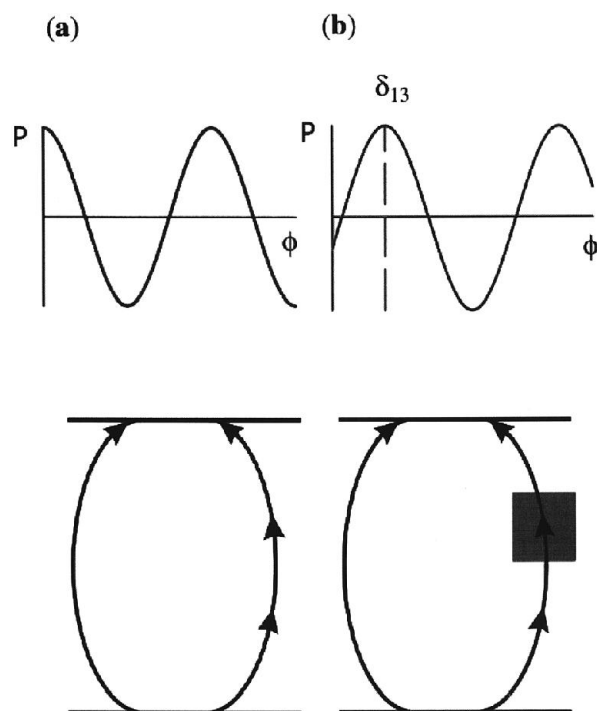


FIGURE 3 Illustration of coherent phase control by one- and three-photon excitation in the cases of (a) no intermediate resonances and (b) a quasi-bound state at the two-photon level that introduces a channel phase shift of δ_{13} . [Reproduced with permission from Fiss, J. A., Khachatryan, A., Truhins, K., Zhu, L., Gordon, R. J., and Seideman, T. (2000). *Phys. Rev. Lett.* **85**, 2096.]

associated with each independent path connecting the initial and final states. In the most commonly studied scenario, illustrated in Fig. 3a, two independent paths are the absorption of three photons of frequency ω_1 and one photon of frequency $\omega_3 = 3\omega_1$. Denoting the j -photon dipole operator by $D^{(j)}$, the probabilities of independent one- and three-photon transitions from the ground state, $|g\rangle$, to a bound excited eigenstate, $|e\rangle$, are given respectively by

$$P_3 = |\langle g|D^{(1)}|e\rangle|^2 \quad (6)$$

and

$$P_1 = |\langle g|D^{(3)}|e\rangle|^2. \quad (7)$$

In the chemically interesting case that the excited state is a continuum leading asymptotically to a product channel S at total energy E , the one-photon transition probability, integrated over all product scattering angles, $\hat{k}$, is given by

$$P_3^S = \left| \int d^3\hat{k} \langle g|D^{(1)}|E, S, \hat{k}\rangle \right|^2, \quad (8)$$

with an analogous equation for P_1^S . The one- and three-photon dipole operators are given by

$$D^{(1)} = -\mu \cdot \varepsilon \quad (9)$$

and

$$D^{(3)} = \sum_{i,j} \frac{D^{(1)}|i\rangle\langle i|D^{(1)}|j\rangle\langle j|D^{(1)}}{(E - \omega_1)(E - 2\omega_1)}, \quad (10)$$

where μ is the electronic dipole, ε is the electric field, and the sum is over all states of the molecule.

For a single excitation path (i.e., one or three photons), the only possibility for controlling the outcome of the reaction is to select the excited eigenstate by varying E , as is normally done in mode-selective processes. A completely new form of control becomes possible, however, if both excitation paths are simultaneously available. In that case, the reaction probability is

$$P^S = \left| \int d^3\hat{k} \langle g|D^{(1)} + D^{(3)}|E, S, \hat{k}\rangle \right|^2 \quad (11)$$

We assume that the electric field is a plane wave linearly polarized in the x direction,

$$\varepsilon_s(t) = \varepsilon_{s0} \hat{x} e^{i(k_{s,z}z - \omega_s t + \varphi_s)} \quad (12)$$

where $k_{s,z}$ is the wave number, φ_s is an arbitrary phase, and $S = 1$ or 3 . It is essential that there be a definite phase relation between the two laser fields, such that

$$\varphi = \varphi_3 - 3\varphi_1 \quad (13)$$

is constant during an experimental run. Inserting Eqs. (9), (10), (12) and (13) into Eq. (11) and expanding the square, one obtains for the reaction probability

$$P^S = P_1^S + P_3^S + 2|P_{13}^S| \cos(\delta_{13}^S + \varphi), \quad (14)$$

where the cross term is given by

$$|P_{13}^S| e^{i\delta_{13}^S} = e^{-i\varphi} \int d^3\hat{k} \langle g|D^{(1)}|ES\hat{k}\rangle \langle ES\hat{k}|D^{(3)}|g\rangle. \quad (15)$$

We refer to δ_{13}^S as a *channel phase*, which is a channel-specific property of the continuum.

The relative phase of the lasers, φ , is a new experimental tool. It is evident from Eq. (14) that the yield of each channel varies sinusoidally with φ . More importantly, one may maximize the relative yield of channel S by setting $\varphi = -\delta_{13}^S$. An experimental signature of phase control is a *phase lag* between the yields from any pair of channels,

$$\Delta\delta(S, S') = \delta_{13}^S - \delta_{13}^{S'}. \quad (16)$$

A theoretical calculation of the branching ratio for the reaction

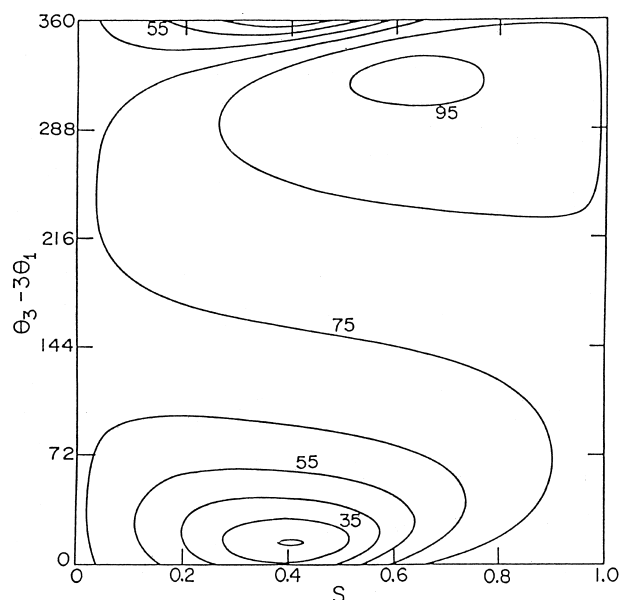
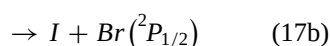
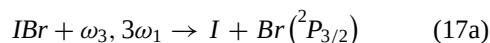


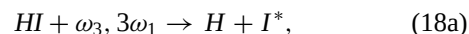
FIGURE 4 Theoretical calculation of the coherent phase control of the photodissociation of IBr by one- and three-photon excitation. [Reproduced with permission from Chan, C. K., Brumer, P., and Shapiro, M. (1991). *J. Chem. Phys.* **94**, 2688. Copyright American Institute of Physics.]



as a function of relative laser phase and intensity is given in Fig. 4. The contours show the fraction of $Br(^2P_{1/2})$ produced from an initial rovibrational level with quantum numbers $v = 0$, $J = 42$, averaged over initial M_J , with a photon energy of $\omega_1 = 6635.0 \text{ cm}^{-1}$. The abscissa is the dimensionless ratio $x^2/(1+x^2)$, where $x = |\epsilon_1|^3/(\bar{\epsilon}|\epsilon_3|)$, and $\bar{\epsilon}$ is defined as a single unit of electric field.

An apparatus used for coherent control experiments is depicted in Fig. 5. A molecular beam intersects the two laser beams in a plane lying between the repeller and extractor electrodes of a time-of-flight mass spectrometer. A uv laser beam of frequency ω_1 is focused into a cell containing a rare gas such as xenon. Third harmonic generation (THG) by the rare gas produces coherent vacuum ultraviolet (vuv) radiation of frequency ω_3 with a definite relative phase, φ , defined by Eq. (13). The two laser beams enter a phase-tuning cell containing a transparent gas such as hydrogen. Because the difference between the indices of refraction at ω_1 and ω_3 is proportional to the pressure of the phase-tuning gas, an increase in the gas pressure produces a linear increase of φ . Ions produced in the reaction region are repelled into a field-free flight tube and detected by a microchannel plate (MCP).

Typical experimental results are shown in Fig. 6 for the reaction



The excited iodine atom produced in reaction (18a) absorbs one or two photons to yield the I^+ ion. The Xe pressure in the third harmonic cell is adjusted so that the one- and three-photon signals are approximately equal. Variation of the H_2 pressure in the phase-tuning cell produces the sinusoidal variation of the ion signals shown in Fig. 6. Evident in this figure is a phase lag of 150° between the two products, HI^+ and I . Also shown is modulation of the signal produced by photoionization of H_2S , which provides a reference phase for the HI^+ and I^+ signals.

Coherent phase control has been used to populate both bound and continuum eigenstates. Bound-to-bound state control has been demonstrated for many molecules, including HCl, CO, NH_3 , CH_3I , $N(CH_3)_3$, $N(C_2H_5)_3$, $(CH_3)_2N_2H_2$, and $c\text{-}C_8H_8$. Bound-to-continuum control has been achieved for the photoionization of Hg, HI, DI, H_2S , and D_2S , and for the photodissociation of HI, DI, and CH_3I . In all of these studies, the use of one- vs three-photon excitation ensures that the parity changes for the two paths are the same. If the parities for the two paths are not equal, as for one- vs two-photon excitation, the average over scattering angles in Eq. (15) causes the cross term to vanish. In this case the differential cross section (i.e., the distribution of recoil angles) may still be controlled. One- vs two-photon control of angular distributions has been demonstrated for the photoionization of Rb and NO and for the photodissociation of HD^+ . This method has also been used to control the direction of an electric current in a GaAs/AlGaAs quantum well and in an amorphous GaAs semiconductor.

It is possible to design multipath control schemes in which the laser phase cancels out of the interference term. One possibility is a “diamond” path configuration, $\omega_1 + \omega_2$ vs $\omega_2 + \omega_1$, with a resonance near ω_1 contributing a phase to the first path and a resonance near ω_2 contributing a phase to the second path. As before, the total probability is the square of the sum of the amplitudes for each path, but here the phases of the two laser beams appear in both paths and cancel in the cross term. In this case the control parameters are the laser frequencies, which determine the detuning from the resonances. This technique was used by Daniel Elliott and coworkers to control the differential cross sections for the ionization of Ba and NO.

Another example of phase-insensitive control utilizes the “lambda” scheme depicted in Fig. 7. In this case a strong coupling (ω_2) field mixes an excited state with

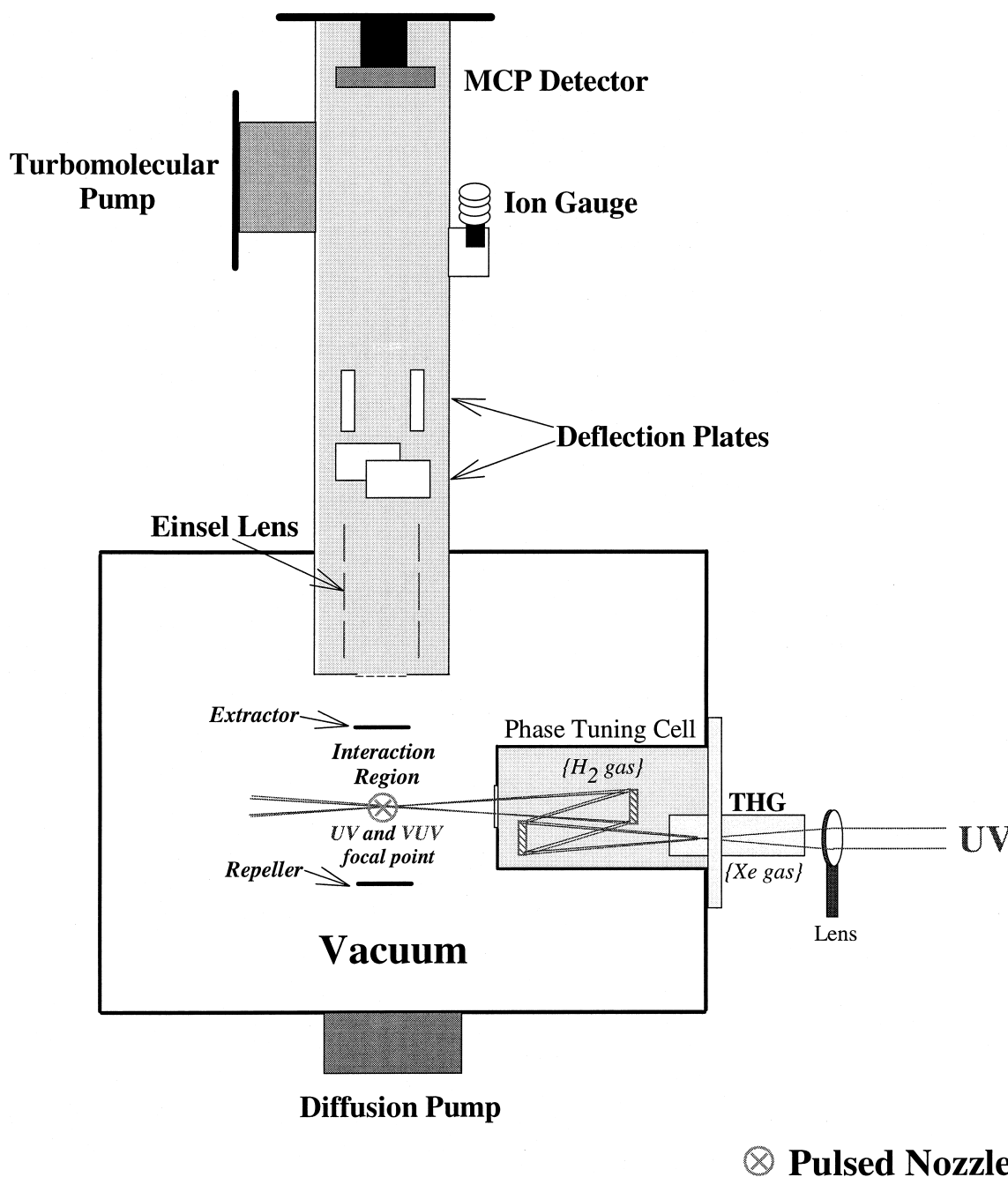


FIGURE 5 Apparatus used for coherent phase control.

the continuum, so that the two paths are $2\omega_1$ and $2\omega_1 - \omega_2 + \omega_2$. (The second path may be viewed as excitation of the continuum by $2\omega_1$ followed by emission and reabsorption of an ω_2 photon.) In the example shown in Fig. 7, a pulsed dye laser (5 ns pulse width) was used to dress the continuum of Na_2 with the $v = 35$, $J = 38$ and $v = 35$, $J = 36$ resonances of the $A^1\Sigma_u/{}^3\Pi_u$ manifold. A second dye laser (ω_1) induced two-photon disso-

ciation of the molecule, and spontaneous emission from $\text{Na}(3p)$ and $\text{Na}(3d)$ was used to monitor the branching ratio of $\text{Na}(3s) + \text{Na}(3p)$ vs $\text{Na}(3s) + \text{Na}(3d)$ fragments as a function of ω_2 detuning. The experimental control of the $\text{Na}(3d)$ product shown in Fig. 8 is in excellent agreement with theory. Comparable results were obtained for the control of $\text{Na}(3p)$, which reached a maximum near $13,317\text{ cm}^{-1}$.

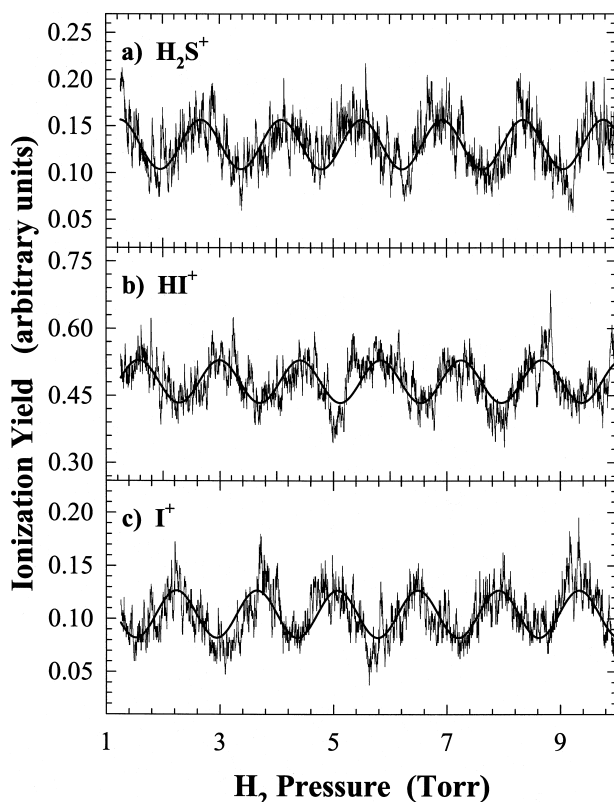


FIGURE 6 Experimental coherent phase control of the photodissociation and photoionization of HI. The three panels show the signals for the ionization of H_2S (top), which is used as phase reference, the ionization of HI (middle), and the dissociation of HI (bottom). [Reproduced with permission from Fiss, J. A., Zhu, L., Gordon, R. J., and Seideman, T. (1999). *Phys. Rev. Lett.* **82**, 65.]

The value of coherent control experiments lies not only in their ability to alter the outcome of a reaction but also in the fundamental information that they provide about molecular properties. In the example of phase-sensitive control, the channel phase reveals information about couplings between continuum states that is not readily obtained by other methods. Examination of Eq. (15) reveals two possible sources of the channel phase—namely, the phase of the three-photon dipole operator $D^{(3)}$, and that of the continuum function, $|ES\hat{k}\rangle$. The former is complex if there exists a metastable state at an energy of ω_1 or $2\omega_1$, which contributes a phase to only one of the paths, as illustrated in Fig. 3b. In this case the channel phase equals the Breit–Wigner phase of the intermediate resonance (modulo π),

$$\delta = -\cot \epsilon, \quad (19)$$

where the ϵ is the reduced energy,

$$\epsilon = 2(E - E_{\text{res}} - \Delta)/\Gamma, \quad (20)$$

and E_{res} , Δ , and Γ are, respectively, the unperturbed resonance center, the shift of the resonance, and its width. An example of this effect is illustrated in Figs. 9 and 10 for the photoionization of HI. The potential energy curves in Fig. 9 display the $b^3\Pi_1$ Rydberg state at approximately two-thirds of the ionization threshold. This quasibound state is predissociated by the $A^1\Pi$ continuum state. The structure evident in the phase lag spectrum in Fig. 10a is produced by the rotational levels of the $b^3\Pi_1$ state. The same rotational structure is evident in the conventional $2 + 1$ resonance-enhanced multiphoton ionization (REMPI) spectrum (Fig. 10b). The absence of any structure in the single-photon ionization spectra of HI (Fig. 10c) and H_2S (Figure 10d) confirms that the phase lag is produced by an intermediate resonance of HI.

The other source of a channel phase is the complex continuum wave function at the final energy E . At first it would appear from Eq. (15) that the phase of $|ES\hat{k}\rangle$ should cancel in the cross term. This conclusion is valid if the product continuum is not coupled either to some another continuum (i.e., if it is elastic) or to a resonance at energy E . If the continuum is coupled to some other continuum (i.e., if it is inelastic), the product scattering wave function can be expanded as a linear combination of continuum functions,

$$|ES\hat{k}\rangle = c_1|ES_1\hat{k}\rangle + c_2|ES_2\hat{k}\rangle, \quad (21)$$

producing a nonzero channel phase that is only weakly energy dependent. The presence of a resonance at energy E produces an extremum in the energy dependence of $|\delta_{13}^s|$. If the underlying continuum is elastic, $|\delta_{13}^s|$ reaches a maximum on resonance, whereas if it is inelastic $|\delta_{13}^s|$ reaches a minimum on resonance. In the limiting case of an isolated resonance coupled to an elastic continuum, with both direct and resonance-mediated transitions to the medium, the channel phase has a Lorentzian energy dependence,

$$\tan \delta_{13}^s = \frac{2(q^{(1)} - q^{(3)})}{\left[\epsilon - \frac{1}{2}(q^{(1)} + q^{(3)})\right]^2 + \left[4 - \frac{1}{4}(q^{(1)} - q^{(3)})^2\right]}, \quad (22)$$

where $q^{(j)}$ is the j -photon Fano shape parameter.

Examples of the latter two sources of the channel phase are illustrated in Fig. 11. From an independent knowledge that the channel phase for ionization of H_2S is zero (or π), it is deduced that the phase lags $\Delta\delta(I, \text{H}_2\text{S}^+)$ and $\Delta\delta(\text{HI}^+, \text{H}_2\text{S}^+)$ are equal, respectively, to the channel phases (modulo π) for the dissociation (δ_{13}^I) and ionization (δ_{13}^{HI+}) of HI. The nearly flat, nonzero values of δ_{13}^I (triangles in Fig. 11a) is indicative of coupling in the dissociation continuum, whereas the peak in δ_{13}^{HI+} (diamonds)

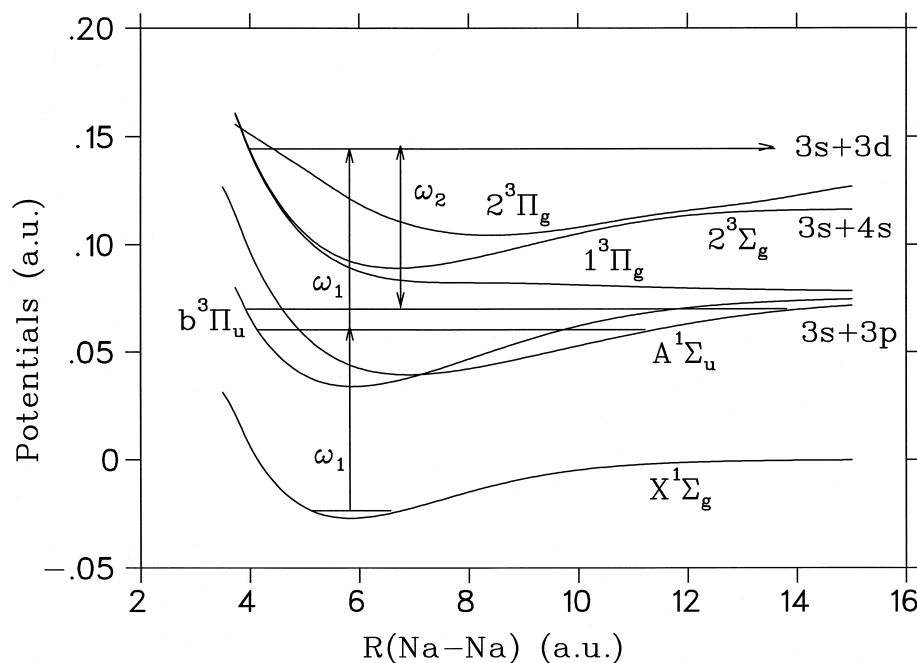


FIGURE 7 Potential energy curves for Na_2 , showing the excitation scheme for incoherent phase control of the photodissociation of the molecule. [Reproduced with permission from Chen, Z., Shapiro, M., and Brumer, P. (1993). *J. Chem. Phys.* **98**, 6843. Copyright American Institute of Physics.]

at $\lambda_1 = 356.2$ nm (the three-photon wavelength) is caused by the $5d(\pi, \delta)$ resonance of HI. This resonance is evident in the one-photon ionization spectrum (Fig. 11b), but is absent in the $2 + 1$ REMPI spectrum (Fig. 11c).

The secondary maximum observed in the phase lag near $\lambda_1 = 355.4$ nm has been attributed to a weak transition to a vibrationally excited Rydberg state not visible in the ionization spectrum. Examples of a minimum in $|\delta_{13}^s|$,

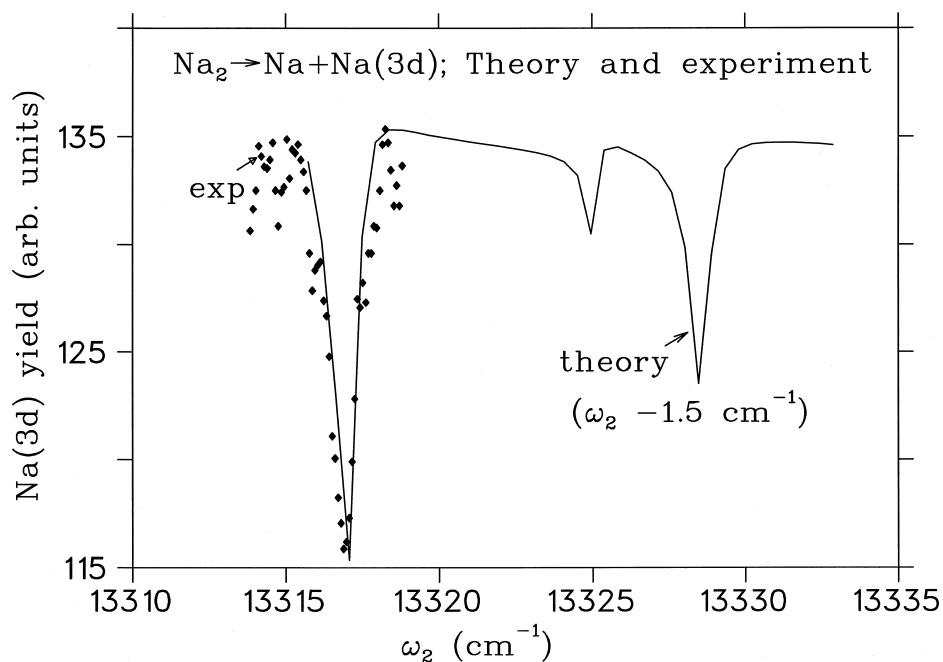


FIGURE 8 Comparison of the experimental and theoretical yields for the reaction $\text{Na}_2 \rightarrow \text{Na}(3s) + \text{Na}(3d)$, obtained by incoherent phase control. [Reproduced with permission from Chen, Z., Shapiro, M., and Brumer, P. (1993). *J. Chem. Phys.* **98**, 6843. Copyright American Institute of Physics.]

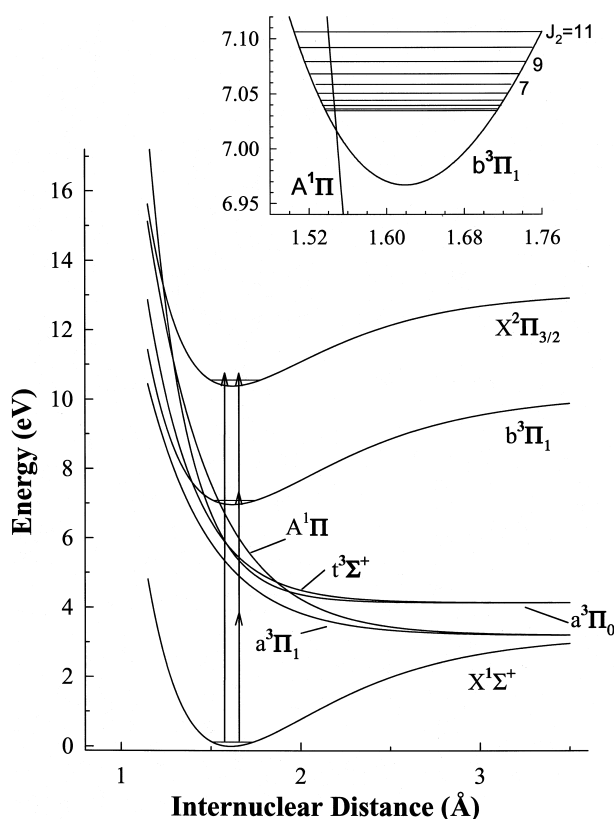


FIGURE 9 Potential energy curves of HI. The insert shows the rotational structure of the $b^3\Pi_1$ Rydberg state, which is predissociated by the $A^1\Pi$ valence state. [Reproduced with permission from Fiss, J. A., Khachatryan, A., Truhins, K., Zhu, L., Gordon, R. J., and Seideman, T. (2000). *Phys. Rev. Lett.* **85**, 2096.]

caused by inelastic coupling to a continuum, have also been observed.

IV. WAVE PACKET CONTROL

A. Introduction

Recent progress in laser technology has led to the widespread use of ultrafast lasers with pulse widths shorter than the vibrational periods of most chemical bonds. A localized state, called a nuclear wave packet, is created on a potential surface by exciting a molecule with ultrashort pulses of radiation. The time-evolution of such wave packets can be directly utilized to observe the transition states of chemical reactions. This development is one of the major accomplishments of femtosecond chemistry.

To understand how wave packets are created by ultrashort pulses, consider a molecule interacting with a pulsed laser field $\epsilon(t)$. Figure 12 shows the time evolution of a

wave packet in a two-electronic state model. In this diagram, X_{g0} represents the lowest vibrational state of the ground electronic state, $X_e^{(1)}(0)$ is the excited-state wave packet created from X_{g0} by an optical excitation, $X_e^{(1)}(\tau)$ is the wave packet at time τ , and $X_g^{(2)}(\tau)$ represents the ground-state wave packet created by stimulated emission from $X_e^{(1)}(\tau)$. Superscripts of X denote the order of the photon-molecule interactions that are used in calculating these wave packets.

The total Hamiltonian $H(t)$ is given within the semiclassical treatment of the molecule-laser field interaction as

$$H(t) = H_0 - \boldsymbol{\mu} \cdot \boldsymbol{\epsilon}(t). \quad (23)$$

Here H_0 is the molecular Hamiltonian, and $\boldsymbol{\mu} \cdot \boldsymbol{\epsilon}(t)$ is the interaction between the molecule and the laser field in the dipole approximation, where $\boldsymbol{\mu}$ is the transition dipole moment of the molecule. Time evolution of the system is determined by the time-dependent Schrödinger equation,

$$i\hbar \frac{\partial}{\partial t} |\Psi(t)\rangle = H(t) |\Psi(t)\rangle. \quad (24)$$

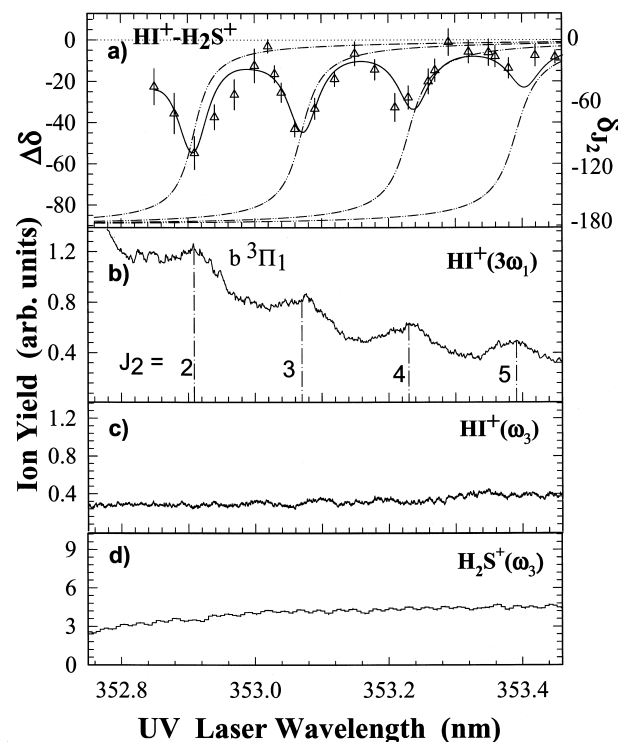


FIGURE 10 The effect of an intermediate resonance on the phase lag for the photoionization of HI. Shown are (a) the phase lag of HI^+ relative to H_2S^+ , (b) the three-photon ionization spectrum of HI, and the one-photon ionization spectra of (c) HI and (d) H_2S . [Reproduced with permission from Fiss, J. A., Khachatryan, A., Truhins, K., Zhu, L., Gordon, R. J., and Seideman, T. (2000). *Phys. Rev. Lett.* **85**, 2096.]

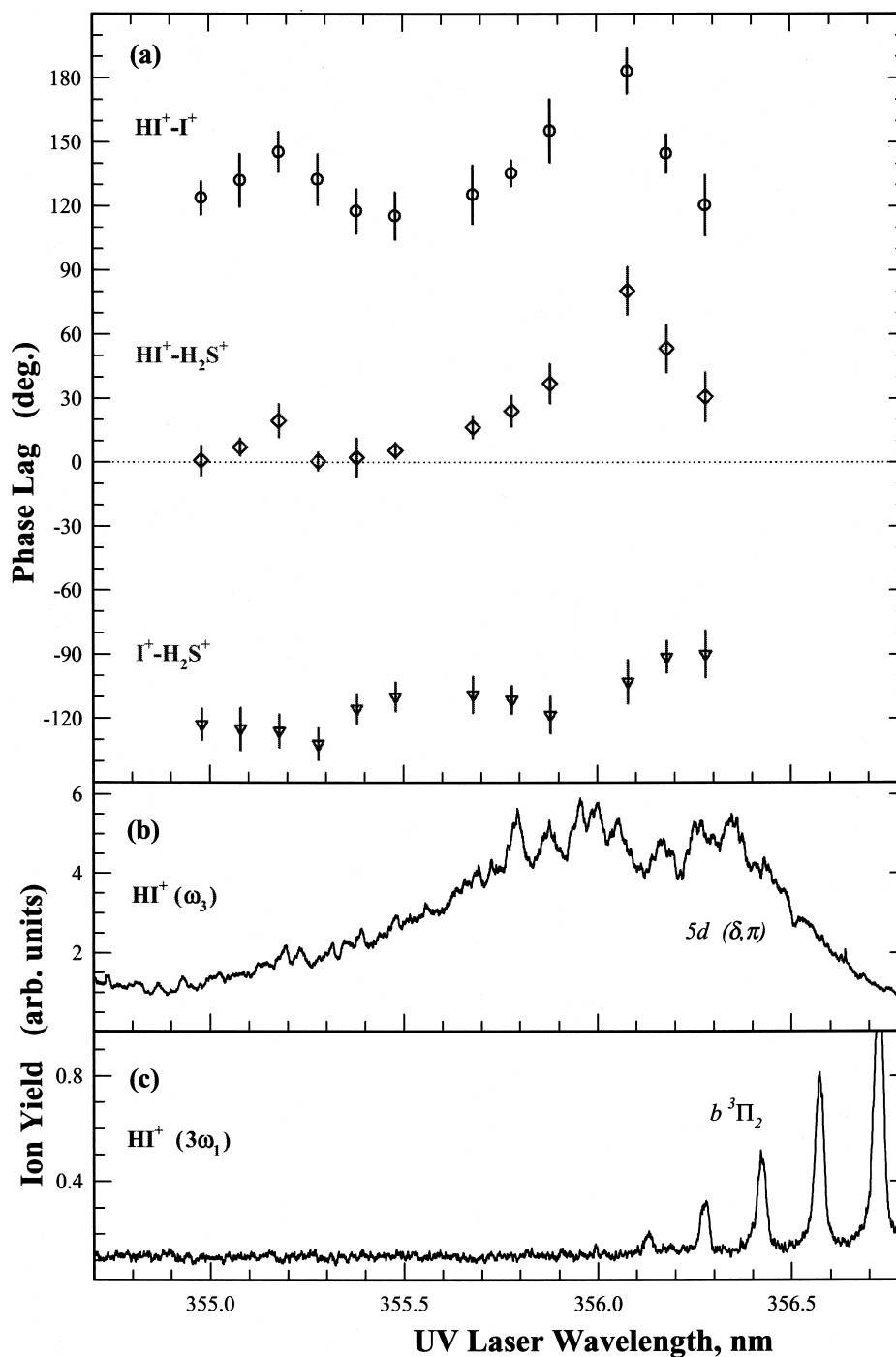


FIGURE 11 Phase lags for the photodissociation and photoionization of HI in the vicinity of the $5d(\pi, \delta)$ resonance. [From Fiss, J. A., Khachatryan, A., Zhu, L., Gordon, R. J., and Seidman, T. (1999). *Disc. Faraday. Soc.* **113**, 61. Reproduced by permission of the Royal Society of Chemistry.]

The solution of this equation $\Psi(t)$ is with the initial condition $\Psi(0)$ at $t = 0$.

Within the Born–Oppenheimer approximation, the initial wave function is expressed as $\Psi(0) = \Phi_g(r, R)X_{gv}(R)$,

where $\Phi_g(r, R)$ denotes the electronic wave function and $X_{gv}(R)$ denotes the nuclear wave function. Here r and R are the coordinates of the electrons and nuclei, respectively. In the case of a weak field in which the

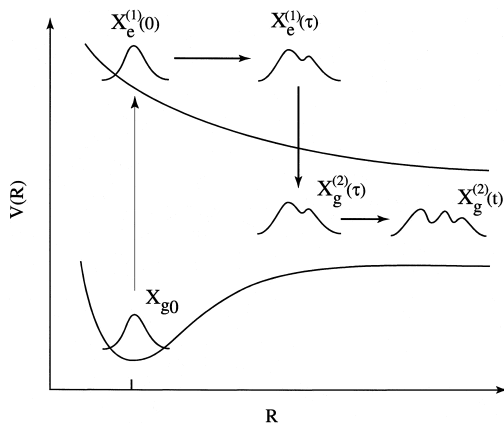


FIGURE 12 Time evolution of a wave packet, based on a perturbative treatment of a two-electronic state model.

population change is negligibly small, first-order time-dependent perturbation theory is sufficient for evaluating the time evolution of the molecular system. The nuclear wave packet $X_e^{(1)}(t)$ created on an electronic potential energy surface e from the lowest ground state, $\Phi_g(r, R)X_{g0}(R)$ in the vibrationless (low temperature) limit, is expressed, as shown in Fig. 12, as

$$|X_e^{(1)}(t)\rangle = \frac{i}{\hbar} \int_0^t dt_1 \exp\left[-\frac{iH_e}{\hbar}(t-t_1)\right] \mu_{eg}(R) \times \exp\left[-\frac{iH_g t_1}{\hbar}\right] |X_{g0}(R)\rangle \varepsilon(t_1), \quad (25)$$

where H_e , the nuclear Hamiltonian of the electronic state e , is given by

$$H_e = T_R + V_e(R). \quad (26)$$

Here T_R is the kinetic energy of the nuclei and $V_e(R)$ is the potential energy. (The ground state Hamiltonian H_g , is given by an equivalent expression with V_e replaced by V_g .) In Eq. (25), $\mu_{eg}(R)$ is the electronic transition moment at the nuclear configuration R . Dephasing effects have been omitted in this equation. They can be taken into account by introducing an effective Hamiltonian, $H_{eff} = H_e - i\Gamma_{eg}$, in which $\Gamma_{eg} = \frac{1}{2}(\Gamma_{gg} + \Gamma_{ee}) + \Gamma'_{eg}$ is a dephasing constant. Here, Γ_{gg} and Γ_{ee} are the decay widths of the ground and excited states, respectively, and Γ'_{eg} is a pure dephasing constant produced by elastic collisions between the molecule and its reservoir.

Let the laser field be expressed as $\varepsilon(t) = \varepsilon_0(t) \sin(\omega_L t)$, where $\varepsilon_0(t)$ is the pulse envelope function, including polarization, and ω_L is the central frequency. For simplicity, consider a δ function excitation. In the rotating wave approximation, the field is expressed as $\varepsilon(t) = \frac{1}{2}\varepsilon_0\delta(t)\exp(-i\omega_L t)$ with field strength ε_0 . Integration over t_1 in Eq. (25) gives

$$|X_e^{(1)}(t)\rangle = \exp\left[-\frac{iH_e t}{\hbar}\right] |X_e^{(1)}(0)\rangle, \quad (27)$$

where $X_e^{(1)}(0)$ is the nuclear wave packet created on the electronically excited state e just after the delta function excitation, given by

$$|X_e^{(1)}(0)\rangle = \frac{i\varepsilon_0}{2\hbar} \mu_{eg}(R) |X_{g0}\rangle. \quad (28)$$

The time evolution of the wave packet given by Eq. (27) can be evaluated by an eigenfunction expansion or by a split-operator technique. With the latter technique, the wave packet $X_e^{(1)}(t + \delta t)$ after a small increment of time δt can be expanded approximately as

$$|X_e^{(1)}(t + \delta t)\rangle = \exp\left[-\frac{iV_e(R)\delta t}{2\hbar}\right] \exp\left[-\frac{iT_R\delta t}{\hbar}\right] \times \exp\left[-\frac{iV_e(R)\delta t}{2\hbar}\right] |X_e^{(1)}(t)\rangle. \quad (29)$$

This expansion is valid to second order with respect to δt . This is a convenient and practical method for computing the propagation of a wave packet. The computation consists of multiplying $|X_e^{(1)}(t)\rangle$ by three exponential operators. In the first step, the wave packet at time t in the coordinate representation is simply multiplied by the first exponential operator, because this operator is also expressed in coordinate space. In the second step, the wave packet is transformed into momentum space by a fast Fourier transform. The result is then multiplied by the middle exponential function containing the kinetic energy operator. In the third step, the wave packet is transformed back into coordinate space and multiplied by the remaining exponential operator, which again contains the potential.

Evolution of the wave packet on the excited state potential energy surface is described by Eq. (27). In the case of a bound potential energy surface, the wave packets are initially localized in the Franck–Condon region but eventually become delocalized because of vibrational mode-mixing processes produced by anharmonicities or because of kinetic couplings. In contrast, if the excited state is unbound, the wave packet rapidly departs from the Franck–Condon region. In both cases time evolution of the wave packet is observed by applying a second pulse, called a probe pulse, which induces stimulated emission or ionization after a selected time delay. This spectroscopic method, known as the pump–probe technique, is used to study the transition state on a femtosecond time scale. From an analysis of the pump–probe spectrum, information about the excited-state dynamics as well as structural properties of the excited potential energy surface may be obtained.

In the weak field limit, the time evolution of wave packets in pump-probe experiments can be evaluated by second-order time-dependent perturbation theory. The second-order solution of Eq. (24) is expressed as

$$\begin{aligned} |X_g^{(2)}(t)\rangle = & -\frac{1}{\hbar^2} \int_0^t dt_2 \int_0^{t_2} dt_1 \exp\left[-\frac{iH_g}{\hbar}(t-t_2)\right] \\ & \times \mu_{ge}(R) \exp\left[-\frac{iH_e}{\hbar}(t_2-t_1)\right] \mu_{eg}(R) \\ & \times \exp\left[-\frac{iH_g t_1}{\hbar}\right] |X_{g0}(R)\rangle \varepsilon(t_2) \varepsilon(t_1). \quad (30) \end{aligned}$$

If both the pump and probe pulses are assumed to be δ functions, Eq. (30) can be expressed as

$$|X_g^{(2)}(t)\rangle = \exp\left[-\frac{iH_g}{\hbar}(t-\tau)\right] |X_g^{(2)}(\tau)\rangle \quad (31)$$

where τ is the delay time between the pump and probe pulses, and $|X_g^{(2)}(\tau)\rangle$, the ground-state wave packet created just after irradiation by the probe pulse, has the form

$$|X_g^{(2)}(\tau)\rangle = \frac{i\varepsilon_0}{2\hbar} \mu_{ge}(R) \exp\left[-\frac{iH_e\tau}{\hbar}\right] |X_e^{(1)}(0)\rangle. \quad (32)$$

In the discussion so far, instantaneous excitation or de-excitation by a δ function pulse has been assumed to transfer wave packets from one electronic state to another state. For realistic pulses, the wave packets may be obtained by numerically integrating Eqs. (25) and (30).

B. Controlling Wave Packets with Tailored Laser Pulses

1. Perturbative Treatment

An intuitive method for controlling the motion of a wave packet is to use a pair of pump-probe laser pulses, as shown in Fig. 13. This method is called the pump-dump control scenario, in which the probe is a controlling pulse that is used to create a desired product of a chemical reaction. The controlling pulse is applied to the system just at the time when the wave packet on the excited state potential energy surface has propagated to the position of the desired reaction product on the ground state surface. In this scenario the control parameter is the delay time τ . This type of control scheme is sometimes referred to as the Tannor-Rice model.

There are many other variables in addition to τ that may be used to control the reaction products by manipulating the motion of wave packets. These include the time-dependent frequency, amplitude, and phase functions of the laser pulse. The use of tailored laser fields to alter the shape of a wave packet is a very general method for controlling the outcome of a chemical reaction.

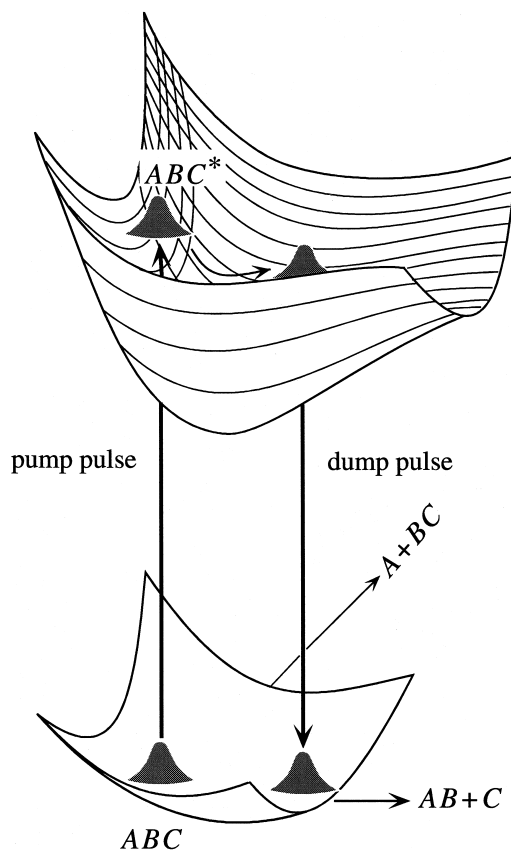


FIGURE 13 A pump-dump control scheme, used to control the branching ratio of the dissociation of a triatomic molecule, ABC.

In order to control a wave packet with tailored laser pulses, we introduce a target operator W , which is a projection operator ($W^2 = W$) that localizes the wave packet at a target position on an electronically excited potential energy surface. The target operator is one of the fundamental quantities in control problems. There are several kinds of target operators, depending on the type of object that is to be controlled. For population control, if a vibronic state X_{ef} is chosen as the target, then its target operator is expressed as $W = |X_{ef}\rangle\langle X_{ef}|$. For wave packet shaping, if a Gaussian wave packet characterized by its average position $\bar{R}$ and average momentum $\bar{P}$ is placed on an electronically excited state e , its target operator is expressed as $W_G = |X_{eG}\rangle\langle X_{eG}|$, with a coordinate representation expressed as $\langle R|W_G|R\rangle = \langle R|X_{eG}\rangle\langle X_{eG}|R\rangle$. Here $\langle R|X_{eG}\rangle$ is given as

$$\langle R|X_{eG}\rangle = (2\pi a^2)^{-\frac{1}{4}} \exp\left[i\frac{\bar{P}}{\hbar}(R - \bar{R}) - \frac{(R - \bar{R})^2}{4a^2}\right], \quad (33)$$

where a , the square root of the variance, is the uncertainty in the position of the wave packet.

Applying W to Eq. (25), we obtain

$$W|X_e^{(1)}(t)\rangle = \frac{1}{\hbar} \int_0^t dt_1 W \exp\left[-\frac{iH_e}{\hbar}(t-t_1)\right] \mu_{eg}(R) \times \exp\left[-\frac{iH_g t_1}{\hbar}\right] |X_{g0}\rangle \varepsilon(t_1). \quad (34)$$

For a weak field, the probability $\langle W(t) \rangle$ of the wave packet existing at the target position at time t is given by

$$\begin{aligned} \langle W(t) \rangle &\approx |W|X_e^{(1)}(t)\rangle|^2 = \langle X_e^{(1)}(t) | W | X_e^{(1)}(t) \rangle \\ &= \frac{1}{\hbar^2} \int_0^{t_f} dt_1 \int_0^{t_f} dt_2 \langle X_{g0} | \exp\left[\frac{iH_g t_2}{\hbar}\right] \mu_{ge}(R) \\ &\quad \times \exp\left[\frac{iH_e}{\hbar}(t-t_2)\right] W \exp\left[-\frac{iH_e}{\hbar}(t-t_1)\right] \mu_{eg}(R) \\ &\quad \times \exp\left[-\frac{iH_g t_1}{\hbar}\right] |X_{g0}\rangle \varepsilon(t_2) \varepsilon(t_1). \end{aligned} \quad (35)$$

Consider the problem of wave packet control in a weak laser field. Here “wave packet control” refers to the creation of a wave packet at a given target position on a specific electronic potential energy surface at a selected time t_f . For this purpose, a variational treatment is introduced. In the weak field limit, the wave packet can be calculated by first-order perturbation theory without the need to solve explicitly the time-dependent Schrödinger equation. In strong fields, where the perturbative treatment breaks down, the time-dependent Schrödinger equation must be explicitly taken into account, as will be discussed in later sections.

In the case of a weak field, the variational method is used to determine the properties of the laser pulses required to reach a specified target. For example, consider the shaping of a Gaussian wave packet in which the target is localized at an average position $\bar{R}$ with an average momentum $\bar{P}$. The target operator is given as W_G . To achieve the desired shape of the wave packet, we define an objective function,

$$J = \langle W_G(t_f) \rangle - \frac{1}{2} \int_0^{t_f} dt \lambda(t) |\varepsilon(t)|^2, \quad (36)$$

where $\langle W_G(t_f) \rangle$ is the expectation value of the wave packet localized near a given Gaussian target. The second term on the right-hand side of Eq. (36) is the constraint on the laser pulses, where $\lambda(t)$ is a time-dependent Lagrange multiplier.

Applying the variational procedure to Eq. (36), we obtain for the optimal control pulse the equation,

$$\int_t^{t_f} dt_1 \langle X_{g0} | W_G^S(t_f; t, t_1) | X_{g0} \rangle \varepsilon(t_1) = \lambda(t) \varepsilon(t), \quad (37)$$

where $W_G^S(t_f; t_2, t_1)$ is a symmetrized operator defined as

$$W_G^S(t_f; t_2, t_1) = W_G(t_f; t_2, t_1) + W_G(t_f; t_1, t_2), \quad (38)$$

with

$$\begin{aligned} W_G(t_f; t_2, t_1) &= \exp\left[\frac{iH_g t_2}{\hbar}\right] \mu_{ge}(R) \exp\left[\frac{iH_e}{\hbar}(t_f - t_2)\right] \\ &\quad \times W_G \exp\left[-\frac{iH_e}{\hbar}(t_f - t_1)\right] \mu_{eg}(R) \\ &\quad \times \exp\left[-\frac{iH_g t_1}{\hbar}\right]. \end{aligned} \quad (39)$$

Because W_G^S in Eq. (39) is a Hermitian operator, the eigenvalues $\lambda(t)$ are real and express the yield of a given target.

This procedure is illustrated by the example of an outgoing wave packet of I_2 on the $B^3\Pi_{0+}$ potential energy surface. The wave packet is assumed to be created from the lowest vibrational level in the ground $X^1\Sigma^+$ state. The potential energy curves for the ground and excited states are shown in Fig. 14. The target is defined as a wave packet on the B surface centered at $\bar{R} = 5.84 \text{ \AA}$, with the center of the outgoing momentum corresponding to a kinetic energy of 0.05 eV. The optimal field $\varepsilon(t)$ is a single pulse with a full width at half-maximum of ~ 225 femtoseconds. The time- and frequency-resolved optimal field is shown in Fig. 15. A Wigner transform of the optimal field $F_w(t, \omega)$, given as

$$F_w(t, \omega) = 2\text{Re} \int_0^\infty dt' \varepsilon^*\left(t + \frac{t'}{2}\right) \varepsilon\left(t - \frac{t'}{2}\right) g(t'), \quad (40)$$

is used. Here $g(t)$ is a window function for smoothing of a spectrum originated from a finite time width. The time- and frequency-resolved spectrum indicates the presence of positive chirp, i.e., a frequency increasing with time. This effect can be seen from the fact that the lower energy components of the continuum wave packet take relatively longer times to reach the target position, and the higher energy components take shorter times.

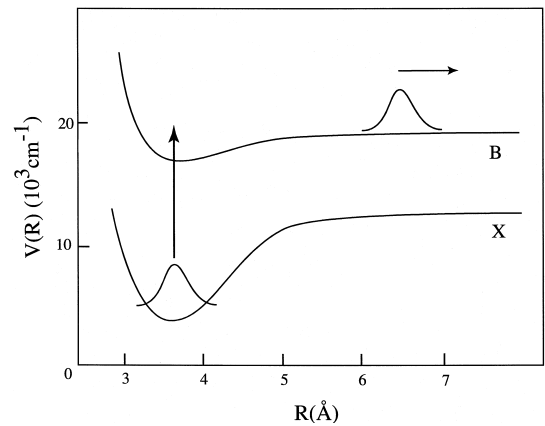


FIGURE 14 Potential energy curves for the ground ($X^1\Sigma^+$) and excited ($B^3\Pi_{0+}$) states of I_2 vapor. Both the initial and outgoing wave packets are shown.

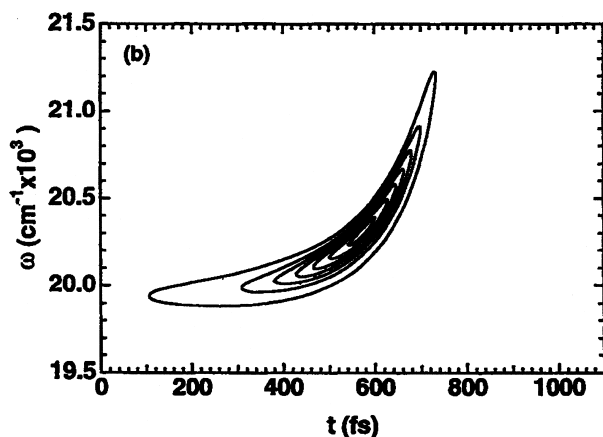


FIGURE 15 Wigner representation of the optimal electric field for I_2 wave packet control. [Reproduced with permission from Krause, Whitnell, J. L., Wilson, R. M., K. R., and Yan, Y. (1993). *J. Chem. Phys.* **99**, 6562. Copyright American Institute of Physics.]

2. Optimal Control

In Section IVB1, a perturbative treatment for wave packet control in a weak field was presented. In this section, a general theory based on an optimal control theory is presented. The resulting expression for laser pulses is applicable to strong as well as weak fields.

The expression for the optimal laser pulse is derived by maximizing the objective function J , defined as

$$J = \hbar \langle W(t_f) \rangle - \frac{1}{2} \int_0^{t_f} \frac{dt}{A(t)} |\varepsilon(t)|^2 + 2\text{Re} \left\{ i \int_0^{t_f} dt \langle \xi(t) | i\hbar \frac{\partial}{\partial t} - H(t) | \psi(t) \rangle \right\}. \quad (41)$$

The first term on the right-hand side of this equation, $\langle W(t_f) \rangle = \langle \psi(t_f) | W | \psi(t_f) \rangle$, is the expectation value of the target operator W at the final time t_f . The second term represents the cost penalty function for the laser pulses with a time-dependent weighting factor $A(t)$. The third term represents the constraint that the wave function $\psi(t)$ should satisfy the time-dependent Schrödinger equation with a given initial condition. Here $\xi(t)$ is the time-dependent Lagrange multiplier.

Carrying out the integration of the third term by parts, the objective function can be rewritten as

$$J = \hbar \langle W(t_f) \rangle - \frac{1}{2} \int_0^{t_f} dt \frac{|\varepsilon(t)|^2}{A(t)} - 2\hbar \text{Re} \left\{ \langle \xi(t) | \psi(t) \rangle \Big|_0^{t_f} \right\} + 2\text{Re} \int_0^{t_f} dt \left\{ \hbar \left\langle \frac{\partial}{\partial t} \xi(t) \right| \psi(t) \right\rangle - i \langle \xi(t) | H(t) | \psi(t) \rangle \right\}. \quad (42)$$

By varying both $\psi(t) \rightarrow \psi(t) + \delta\psi(t)$ and $\varepsilon(t) \rightarrow \varepsilon(t) + \delta\varepsilon(t)$ in the above equation, the objective function J is expressed as $J + \delta J$, where δJ has the form

$$\delta J = \int_0^{t_f} dt \left[2\text{Re} \left\{ i \langle \xi(t) | \mu | \psi(t) \rangle \right\} - \frac{\varepsilon(t)}{A(t)} \right] \delta\varepsilon(t) + 2\text{Re} \int_0^{t_f} dt \left[\hbar \left\langle \frac{\partial}{\partial t} \xi(t) \right| \delta\psi(t) \right] - i \langle \xi(t) | H(t) | \delta\psi(t) \rangle + 2\hbar \text{Re} \{ \langle \psi(t_f) | W | \delta\psi(t_f) \rangle - \langle \xi(t_f) | \delta\psi(t_f) \rangle \}. \quad (43)$$

From the optimal condition, $\delta J = 0$, the expression for the optimal laser pulse,

$$\varepsilon(t) = -2A(t)\text{Im} \langle \xi(t) | \mu | \psi(t) \rangle, \quad (44)$$

is obtained. Here $\xi(t)$ satisfies the time-dependent Schrödinger equation,

$$i\hbar \frac{\partial}{\partial t} |\xi(t)\rangle = H(t) |\xi(t)\rangle, \quad (45)$$

with the final condition at $t = t_f$,

$$|\xi(t_f)\rangle = W |\psi(t_f)\rangle. \quad (46)$$

The optimal pulse can be obtained by solving the time-dependent Schrödinger equation iteratively with initial and final boundary conditions. First, assuming an analytical form for $\varepsilon(t)$, the time-dependent Schrödinger equation is solved to obtain $\psi(t)$ by forward propagation of the molecular system. Second, solving Eq. (45) with the same form of $\varepsilon(t)$ as before, but with the final condition, Eq. (46), the backward propagated wave function $\xi(t)$ can be obtained. A new form of the laser field $\varepsilon(t)$ can then be constructed by substituting these two wave functions, $\psi(t)$ and $\xi(t)$, into Eq. (44). These procedures are repeated until convergence is reached. This is a general procedure for obtaining optimal pulse shapes, and is called the global optimization method. By using this method, one can obtain the true optimal solution of systems having many local solutions. Convergence problems sometimes arise when global optimization is applied to real reaction systems. Several numerical methods for carrying out global optimization, such as the steepest descent method and a genetic algorithm, have been developed.

Another approach is known as the local optimization method. Here “local” means that maximization of the objective function J is carried out at each time, i.e., locally in time between 0 and t_f . There are several methods for deriving an expression for the optimal laser pulse by local optimization. One is to use the Ricatti expression for a linear time-invariant system in which a differential equation of a function connecting $\psi(t)$ and $\xi(t)$ is solved, instead of directly solving for these two functions. Another method

is to solve an inverse problem for the path in a functional space of J . There are two essential points in the local optimization method. The first is to divide the time interval t_f from the initial time into infinitesimally short time intervals. The second point is to impose the final condition at the end of each time interval, i.e., $|\xi(t)\rangle = W|\psi(t)\rangle$. Following this procedure, the optimized pulse at time t is expressed as

$$\varepsilon(t) = -2A(t) \operatorname{Im}\langle\psi(t)|W\mu|\psi(t)\rangle. \quad (47)$$

The simplest method for obtaining this expression is to substitute Eq. (46) into Eq. (44) after changing t_f to t in Eq. (46). This means that a (virtual) target is set just after each infinitesimally small time increment, and then the virtual target is moved toward the final true target position. The necessary condition for local optimization is therefore the assurance of an increase in the population of the target state. The merit of local optimization is that only one-sided propagation, i.e., forward or backward propagation, is needed. Its algorithm is, therefore, quite simple. Once the initial condition is specified, the time-dependent Schrödinger equation with a seed pulse as its initial pulse form can be solved to obtain a wave function after an infinitesimally increased propagation time. Next, substituting the resulting wave function into Eq. (47), an expression for the pulse at the increased time is obtained. With this pulse form, the time-dependent Schrödinger equation is solved again. This cycle of calculation described above is repeated until convergence is reached. This is a form of feedback control, because the wave function and laser pulse are related by Eq. (47). Because the local optimization method described above is nonperturbative, this method can be applied to wave packet control in intense fields. In such an intense field case, Eq. (47) can be used by letting the time increment become smaller and smaller.

As an example of optimal control, we consider the local control of a ring-puckering isomerization such as that of trimethyleneimine. The coordinate of the puckering motion q is defined as the displacement of the line joining the carbon and nitrogen atoms. The adiabatic potential energy expressed as a function of q is a double minimum potential. Figure 16 shows the adiabatic potential energy function together with several vibrational eigenfunctions. A linear dipole moment with respect to q was assumed. The time evolution of the probability density of the wave packet, $|\langle q|\psi(t)\rangle|^2$, produced by the locally optimized laser field is shown in Fig. 17. Starting from isomer A, the wave packet is almost completely transferred to the well of isomer B within 10 ps. Figure 18 shows the time variation of the locally optimized electric field. By analyzing the electric field with the help of a window Fourier transform, the optimized field may be regarded as four

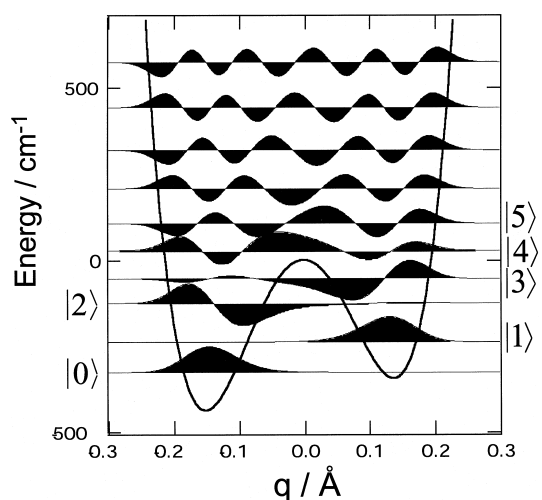


FIGURE 16 Adiabatic potential energy of trimethyleneimine as a function of the puckering coordinate q . The vibrational eigenfunctions are superimposed on the potential energy curve. [Reproduced with permission from Sugawara, M., and Fujimura, Y. (1994). *J. Chem. Phys.* **100**, 5646. Copyright American Institute of Physics.]

successive pulses with carrier frequencies that correspond to transition frequencies of the molecular eigenstates.

An advantage of the local control method described above is that it can be applied to wave packet propagation starting from an initial, nonstationary state, in contrast to ordinary wave packet control, which begins with the initial condition of a stationary state. An example where starting from such an initial condition is useful is the control of a localized state of a double-well potential. In this case, by propagating the final-state wave packet backward to the initial state, pulses that are optimized for forward

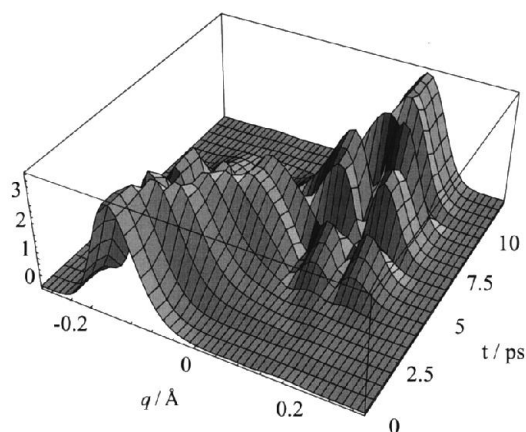


FIGURE 17 Wave packet dynamics of trimethyleneimine produced by an optimized laser field. [Reproduced with permission from Sugawara, M., and Fujimura, Y. (1994). *J. Chem. Phys.* **100**, 5646. Copyright, American Institute of Physics.]

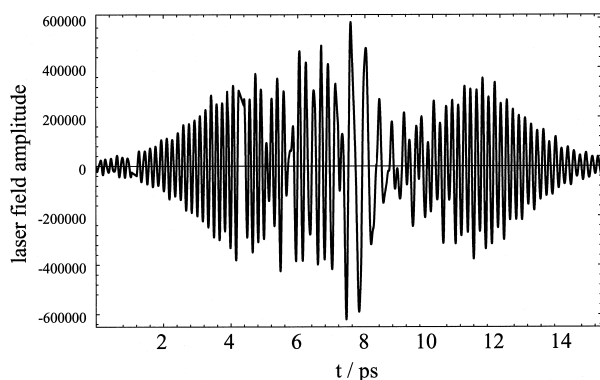


FIGURE 18 Time variation of the optimized electric field. [Reproduced with permission from Sugawara, M., and Fujimura, Y. (1994). *J. Chem. Phys.* **100**, 5646. Copyright American Institute of Physics.]

propagation can be constructed by time reversal, because the time-dependent Schrödinger equation is unitary in the case of no dissipation.

So far we have treated only the case of wave packets constructed from pure states. Consider now the control of a molecular system in a mixed state in which the initial states are distributed at a finite temperature. The time evolution of the system density operator $\rho(t)$ is determined by the Liouville equation,

$$i\hbar \frac{\partial}{\partial t} \rho(t) = L(t)\rho(t), \quad (48)$$

where the Liouville operator $L(t)$ is given by

$$L(t)\rho(t) = H(t)\rho(t) - \rho(t)H(t). \quad (49)$$

It is convenient to introduce a Liouville space, or double space, that is a direct product of cap and tilde spaces. In Liouville space, operators are considered to be vectors and Hilbert-space commutators are considered to be operators. Equation (48) is then expressed as

$$i\hbar \frac{\partial}{\partial t} |\rho(t)\rangle\rangle = L(t)|\rho(t)\rangle\rangle, \quad (50)$$

where $|\rho(t)\rangle\rangle$ is a vector, and

$$L(t) = L_0 - M\varepsilon(t) \quad (51)$$

is an operator in Liouville space. Here $L_0 = \hat{H}_0 - \tilde{H}_0$, and $\hat{H}_0$ and $\tilde{H}_0$ are molecular Hamiltonians in the cap and tilde spaces, respectively. Similarly, $M = \hat{\mu} - \tilde{\mu}$. In the Liouville representation, the objective function is rewritten as

$$J = \hbar \langle\langle W_G | \rho(t_f) \rangle\rangle - \frac{1}{2} \int_0^{t_f} dt \frac{|\varepsilon(t)|^2}{A(t)} + 2\text{Re} i \int_0^{t_f} dt_1 \times \left\{ \langle\langle \Xi(t) | i\hbar \frac{\partial}{\partial t} \rho(t) \rangle\rangle - \langle\langle \Xi(t) | L(t) | \rho(t) \rangle\rangle \right\}. \quad (52)$$

Equation (52) has the same structure as that of Eq. (41). An expression for the optimal control pulse in a mixed case can therefore be obtained as

$$\varepsilon(t) = -2A(t)\text{Im} \langle\langle \Xi(t) | \hat{\mu} | \rho(t) \rangle\rangle, \quad (53)$$

where the time-dependent multiplier $\Xi(t)$ satisfies the Liouville equation,

$$i\hbar \frac{\partial}{\partial t} |\Xi(t)\rangle\rangle = L^\dagger(t) |\Xi(t)\rangle\rangle, \quad (54)$$

with the final condition $|\Xi(t_f)\rangle\rangle = |W\rangle\rangle$.

A fundamental limitation to coherent population control is that it is impossible to transfer 100% of the population in a mixed state. That is, the maximum value of the population transferred cannot exceed the maximum of the initial population distribution of a system without any dissipative process such as spontaneous emission. This result can be simply verified using the unitary property of the density operator, $\rho(t) = U(t, t_0)\rho(t_0)U^\dagger(t, t_0)$, where $\rho(t_0)$ is the diagonalized density operator at $t = t_0$, $U(t, t_0)$ is the time-evolution operator given by

$$U(t, t_0) = \hat{T} \exp \left[-\frac{i}{\hbar} \int_{t_0}^t dt' V_I(t') \right], \quad (55)$$

$\hat{T}$ is a time-ordering operator, and $V_I(t')$ is the interaction between the molecules and the controlling pulses in the interaction representation. The eigenvalues of $\rho(t)$ are thus invariant with respect to unitary transformation. The population of a target state $|k\rangle$ at time t , $\langle k | \rho(t) | k \rangle$, satisfies the condition that the minimum eigenvalue of $\rho(t_0) \leq \langle k | \rho(t) | k \rangle \leq$ the maximum eigenvalue of $\rho(t_0)$. That is, the maximum population in a target state at time t_f is equal to the maximum eigenvalue of $\rho(t_0)$. Therefore, in the mixed state case, one must choose a target operator appropriate for this restriction.

3. Experimental Examples of Wave Packet Control

The key technological advance that has made optical pulse shaping widely available is the pulse modulator depicted in Fig. 19. For a Gaussian laser pulse the product (full-width at half-maximum) of duration τ and radial frequency bandwidth $\delta\omega$ is 0.44. (For a sech^2 pulse, the product is 0.32.) For such transformed-limited pulses the group velocity is the same for all frequencies. The properties of a laser pulse can be tailored by dispersing the pulse, filtering the frequency components, and finally reconstituting the modified pulse. This method is illustrated in Figure 19, where grating G_1 is placed at the focal point of lens L_1 . A multipixel spatial light modulator (SLM) placed in the Fourier plane is programmed to alter the

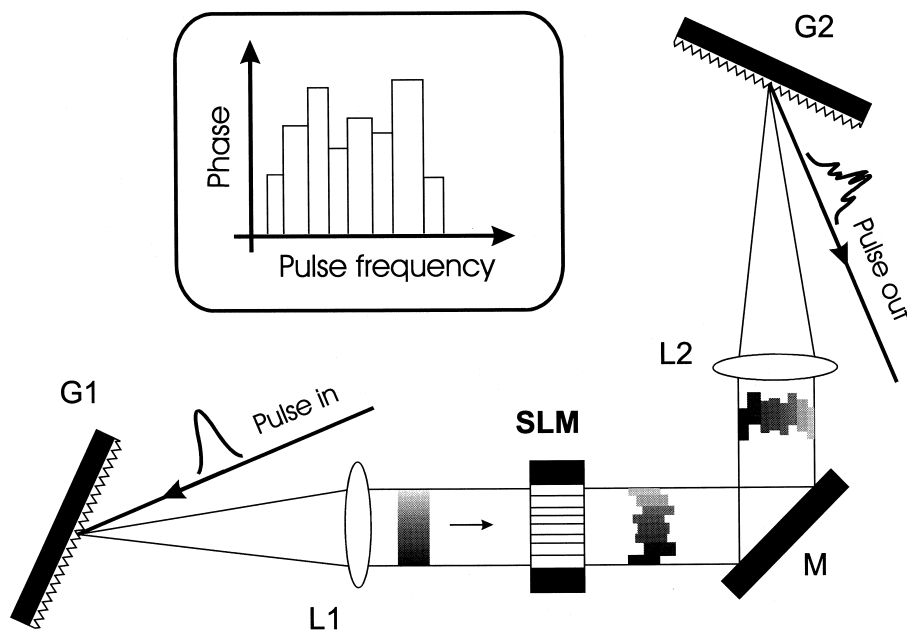


FIGURE 19 Illustration of pulse shaping using a liquid crystal spatial modulator. [Uberna, R., Amitay, Z., Loomis, R. A., and Leone, S. R. (1999). *Disc. Faraday Society* **113**, 385. Reproduced by permission of the Royal Society of Chemistry.]

phases and/or amplitudes of the light in each frequency interval. To control both amplitudes and phases, two modulators are used back-to-back. The two types of devices that have been used for this purpose are a liquid crystal spatial light modulator and an acousto-optic modulator. The modified pulse is finally recompressed by lens L_2 and grating G_2 .

The phase and amplitude spectrum of the laser pulse are tailored to create a wave packet with selected properties. The various eigenstates that comprise the wave packet are populated by different frequency components of the laser pulse, each with its specified amplitude and phase. For example, rovibrational wave packets of Li_2 in the $E^1\Sigma_g^+$ state were created, consisting of vibrational levels $v = 12\text{--}16$ and rotational levels $J = 17, 19$. The phases and amplitudes of the pump pulse shown in Fig. 20 were generated with a 128-pixel liquid crystal SLM. The pulse was tailored to optimize the ionization signal at a delay time of 7 ps. The phases used to maximize or minimize the ionization signal are shown by solid and dashed lines, respectively, and the intensities at the eigenfrequencies of the wave packet are indicated by circles.

C. Genetic Algorithms

There are many cases in which the molecular Hamiltonian and the interactions with the photon fields are not completely known. For many isolated polyatomic molecules or for molecules in condensed phases, for ex-

ample, only the initial and final states of the control system are specified; the multidimensional potential energy surfaces and the reaction coordinates connecting the initial and final states cannot be determined either theoretically or experimentally. In addition, there exist experimentally unavoidable uncertainties associated with the controlling fields. The optimal control methods described in previous sections are not well suited for such systems. In such cases, a genetic algorithm (GA), which is a global optimization method, may be employed.

A GA has three fundamental operations: reproduction, crossover, and mutation. By means of these three operations, an ensemble of individuals (i.e., a population) is adapted to fit its environment. In the first operation, groups of individuals that have a higher probability of reproducing as a consequence of their better adaptation to their environment pass their genetic information onto succeeding generations. In the crossover operation, splicing of the parents' genes transfers a mixture of their genetic material to their offspring. Genetic diversity of the offspring cannot be created by crossover. By mutation, however, the genetic material is altered to avoid premature convergence to an undesirable trait.

The GA procedure has an interactive feedback (closed) loop structure without any intervention by the experimentalist. An initial guess of multiple sets of control fields is evaluated for its success in achieving the target. The most successful members of this population are transformed by the three GA operators, and their offspring are evaluated.

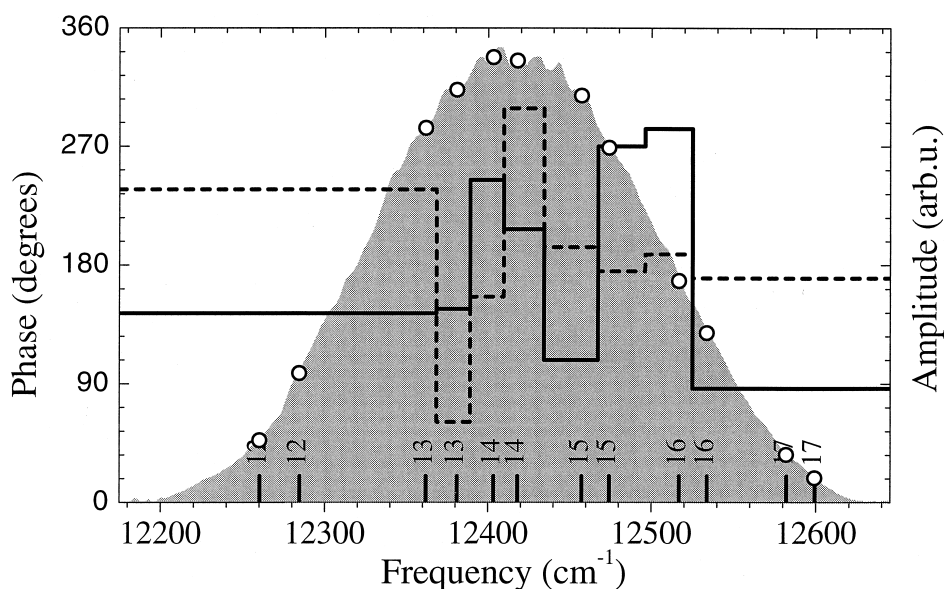


FIGURE 20 Amplitude and phase of a laser pulse optimized either to maximize or minimize the ionization of Li_2 . [Uberna, R., Amitay, Z., Loomis, R. A., and Leone, S. R. (1999). *Disc. Faraday Society* **113**, 385. Reproduced by permission of the Royal Society of Chemistry.]

The entire process is repeated for many generations until convergence is achieved.

An experimental illustration of the GA is shown in Fig. 21. The molecule cyclopentadienyl-iron-dicarbonyl-chloride was irradiated with pulses of 800 nm radiation that were initially 80 ns long before entering the pulse shaper. The phases of the laser pulses were modified with

a liquid crystal SLM and optimized with a GA either to maximize or minimize the ratio of $\text{C}_5\text{H}_5\text{FeCOCl}^+$ to FeCl^+ . Convergence was achieved typically after 100 generations. The optimum yield ratio and the pulse shapes used to achieve them are shown. (A GA was similarly used in the example of Li_2 ionization illustrated in Fig. 20.)

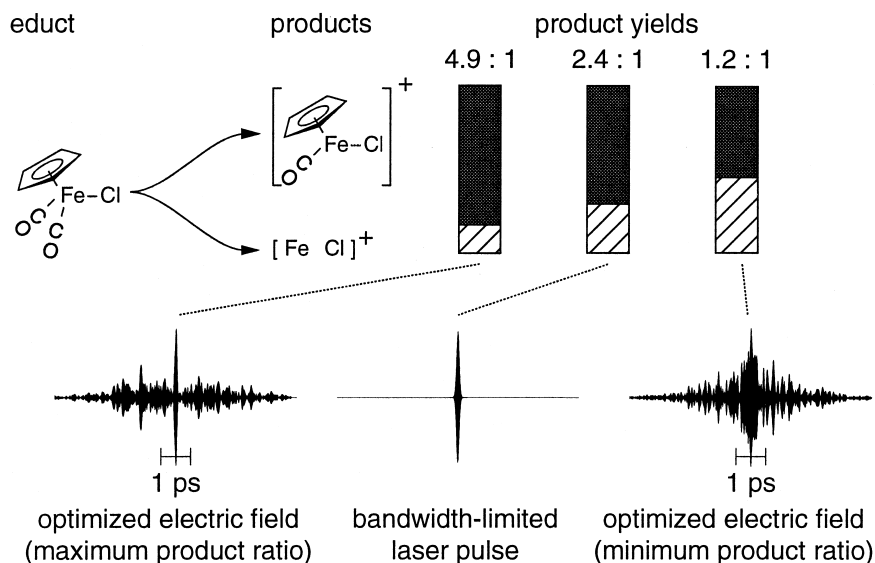


FIGURE 21 Use of the genetic algorithm to control the dissociative ionization of cyclopentadienyl-iron-dicarbonyl chloride. Shown on the bottom are the optimized electric fields generated to either maximize or minimize the ratio of $\text{C}_5\text{H}_5\text{FeCOCl}^+$ to FeCl^+ . [Provided by the courtesy of T. Brixner, and adapted from Assion, A., Baumert, T., Bergt, M., Brixner, T., Kiefer, B., Seyfried, V., Strehler, M., and Gerber, G. (1998). *Science* **282**, 919.]

D. Coherent Population Transfer

So far, we have considered various theoretical treatments of time-dependent wave packets controlled by laser pulses to produce a desired product in a chemical reaction. Another type of problem, based on adiabatic behavior of wave functions, is the transfer of population from one state to another.

Suppose that a laser pulse interacts adiabatically with a molecular system. By the term “adiabatic” is meant that an eigenstate $\psi_\ell(t)$ at time t satisfies the time-independent Schrodinger equation:

$$H(t)\psi_\ell(t) = E_\ell(t)\psi_\ell(t). \quad (56)$$

In the adiabatic limit, t is considered to be a parameter, and $\psi_\ell(t)$ is called an adiabatic state. One of the interesting properties of this limit is that a population can be inverted by evolving the system adiabatically. This process is called adiabatic passage. Population transfer induced by a laser is generally called “coherent population transfer.” For a two-level system, the complete population inversion is produced by a π -pulse or by adiabatic rapid passage.

Population transfer in a three-level system can be achieved by using one laser (known as the “pump laser,” which may be either continuous wave or pulsed) to connect the ground and intermediate levels, and a second laser (the “Stokes laser”) to connect the intermediate and final levels. This method, known as stimulated Raman adiabatic passage or STIRAP, is illustrated in Fig. 22. In this example, the three levels have a Λ -type configuration, where

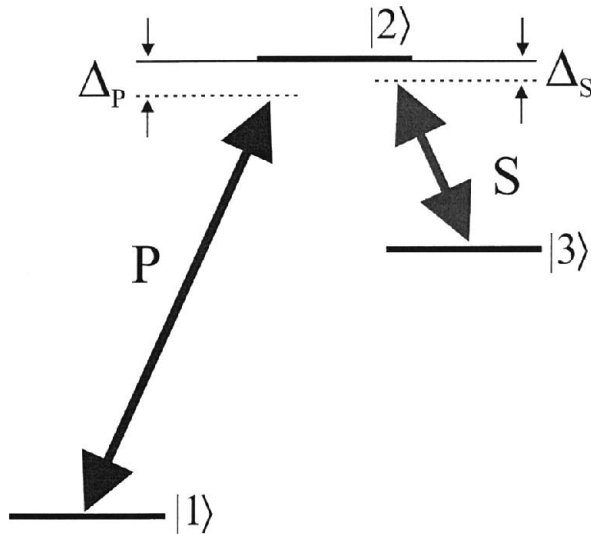


FIGURE 22 Three-level excitation scheme used for STIRAP. [Reproduced with permission from Bergmann, K., Theuer, H., and Shore, B. W. (1998). *Rev. Mod. Phys.* **70**, 1003.]

ϕ_1 , ϕ_2 , and ϕ_3 are the wave functions of the initial, intermediate, and final states, respectively, (denoted by $|1\rangle$, $|2\rangle$, and $|3\rangle$ in Fig. 22), subscripts P and S label the pump and Stokes processes, respectively, and $\Delta_P(\Delta_S)$ denotes the detuning of the pump (Stokes) laser. In the discussion that follows, the two laser fields are assumed to be pulsed.

The interaction matrix elements of the total Hamiltonian are expressed in the three-state model as

$$\frac{\hbar}{2} \begin{bmatrix} -2\Delta & \Omega_P(t) & 0 \\ \Omega_P^*(t) & 0 & \Omega_S(t) \\ 0 & \Omega_S^*(t) & -2\Delta \end{bmatrix}, \quad (57)$$

where $\Omega_P(t)$ is the Rabi frequency of the pump pulse, $\Omega_S(t)$ and is the Rabi frequency of the Stokes pulse. By diagonalizing the determinant of this interaction matrix, the adiabatic states ψ_0 and $\psi_\pm$, with eigenfrequencies ω_0 and $\omega_\pm$, are analytically derived as

$$\psi_0(t) = \cos[\Theta(t)]\phi_1 - \sin[\Theta(t)]\phi_3 \quad (58)$$

with eigenfrequency

$$\omega_0(t) = -\Delta(t), \quad (59)$$

$$\begin{aligned} \psi_+(t) = & \sin[\eta(t)] \sin[\Theta(t)]\phi_1 + \cos[\eta(t)]\phi_2 \\ & + \sin[\eta(t)] \cos[\Theta(t)]\phi_3 \end{aligned} \quad (60)$$

with eigenfrequency

$$\omega_+(t) = -\frac{1}{2} \left[\Delta(t) - \sqrt{|\Omega_P(t)|^2 + |\Omega_S(t)|^2 + \Delta^2(t)} \right], \quad (61)$$

and

$$\begin{aligned} \psi_-(t) = & \cos[\eta(t)] \sin[\Theta(t)]\phi_1 - \sin[\eta(t)]\phi_2 \\ & + \cos[\eta(t)] \cos[\Theta(t)]\phi_3 \end{aligned} \quad (62)$$

with eigenfrequency

$$\omega_-(t) = -\frac{1}{2} \left[\Delta(t) + \sqrt{|\Omega_P(t)|^2 + |\Omega_S(t)|^2 + \Delta^2(t)} \right]. \quad (63)$$

The mixing angle $\Theta(t)$ in Eqs. (58), (60), and (62) is given by

$$\sin[\Theta(t)] = \frac{|\Omega_P(t)|}{\sqrt{|\Omega_P(t)|^2 + |\Omega_S(t)|^2}} \quad (64)$$

$$\cos[\Theta(t)] = \frac{|\Omega_S(t)|}{\sqrt{|\Omega_P(t)|^2 + |\Omega_S(t)|^2}}, \quad (65)$$

and

$$\tan[\eta(t)] = \frac{\sqrt{|\Omega_P(t)|^2 + |\Omega_S(t)|^2}}{\sqrt{|\Omega_P(t)|^2 + |\Omega_S(t)|^2 + \Delta^2(t) + \Delta(t)}}. \quad (66)$$

We note that the lowest adiabatic state, Eq. (58), is expressed in terms of the initial and the final states without the intermediate state. This property implies that the system in the initial state is transferred to the final state adiabatically, with no population in the intermediate resonant state. That is, under the condition for the Rabi frequencies $|\Omega_P(t)| \ll |\Omega_S(t)|$, we can see from Eqs. (64) and (65) that $\psi_0(0) = \phi_1$ and $\psi_0(\infty) = \phi_3$ if the Stokes pulse comes before the pump pulse. The ordering of these pulses in STIRAP is counterintuitive, compared with conventional, stimulated Raman scattering processes. Figure 23 shows the time evolution of the Rabi frequencies, mixing angle, dressed-state eigenvalues, and the population of the initial and final levels. The counterintuitive pulse sequence is evident in Figure 23a. One of the properties of STIRAP

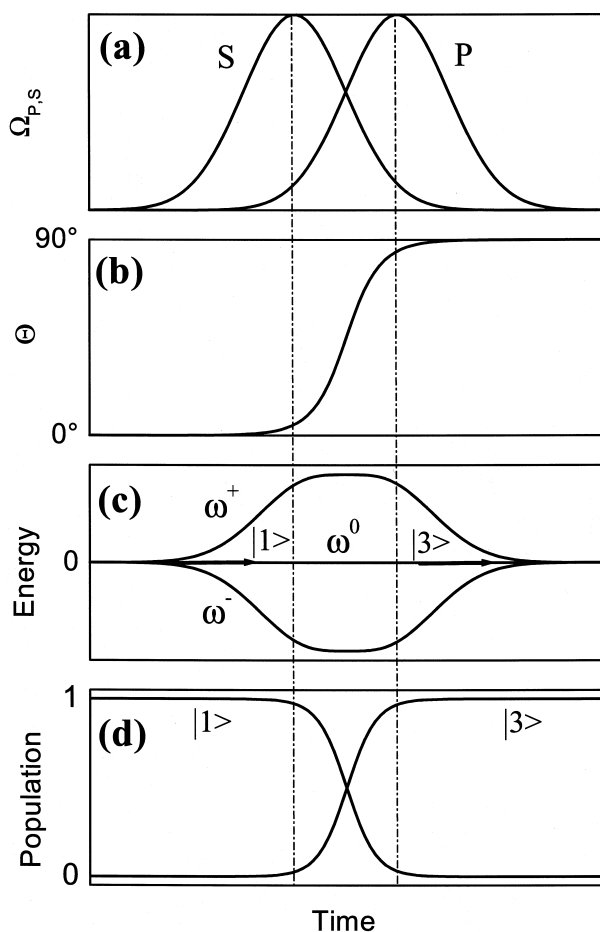


FIGURE 23 Illustration of the STIRAP technique used for coherent population transfer. Shown are the time evolution of (a) the Rabi frequencies of the pump and Stokes lasers, (b) the mixing angle, (c) the dressed-state eigenvalues, and (d) the populations of the initial and final levels. [Reproduced with permission from Bergmann, K., Theuer, H., and Shore, B. W. (1998) *Rev. Mod. Phys.* **70**, 1003.]

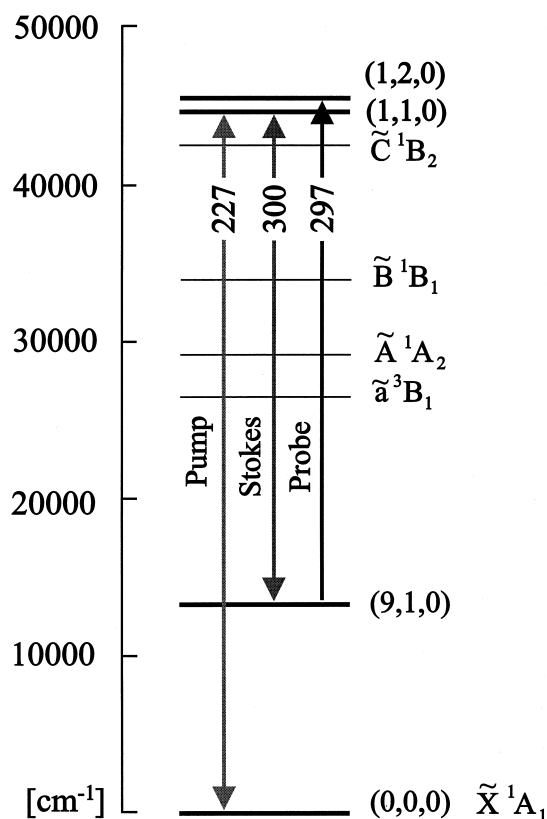


FIGURE 24 Energy levels of SO_2 , showing the pumping scheme used to transfer population from the ground level to the $(9,1,0)$ vibrationally excited level by STIRAP. [Reproduced with permission from Halfmann, T., and Bergmann, K. (1996) *J. Chem. Phys.* **104**, 7068. Copyright American Institute of Physics.]

is its robustness with respect to parameters such as Rabi frequency and time delay between the Stokes and pump pulses. The STIRAP technique can also be applied to a system with more than three levels.

An example of a STIRAP simulation and experiment is illustrated in Figs. 24 and 25 for SO_2 . Figure 24 shows the energies of the laser pulses used to transfer population from the vibrationless level to the $(9,1,0)$ level of the ground electronic state. Figure 25 shows the experimentally measured and numerically simulated fraction of the population transferred to the excited state as a function of the time delay between the pump and Stokes pulses. The greater efficiency of a counterintuitive pulse sequence is evident.

V. CONTROL OF EXTERNAL DEGREES OF FREEDOM

In the discussion so far, emphasis has been placed on controlling the internal degrees of freedom of an atom or

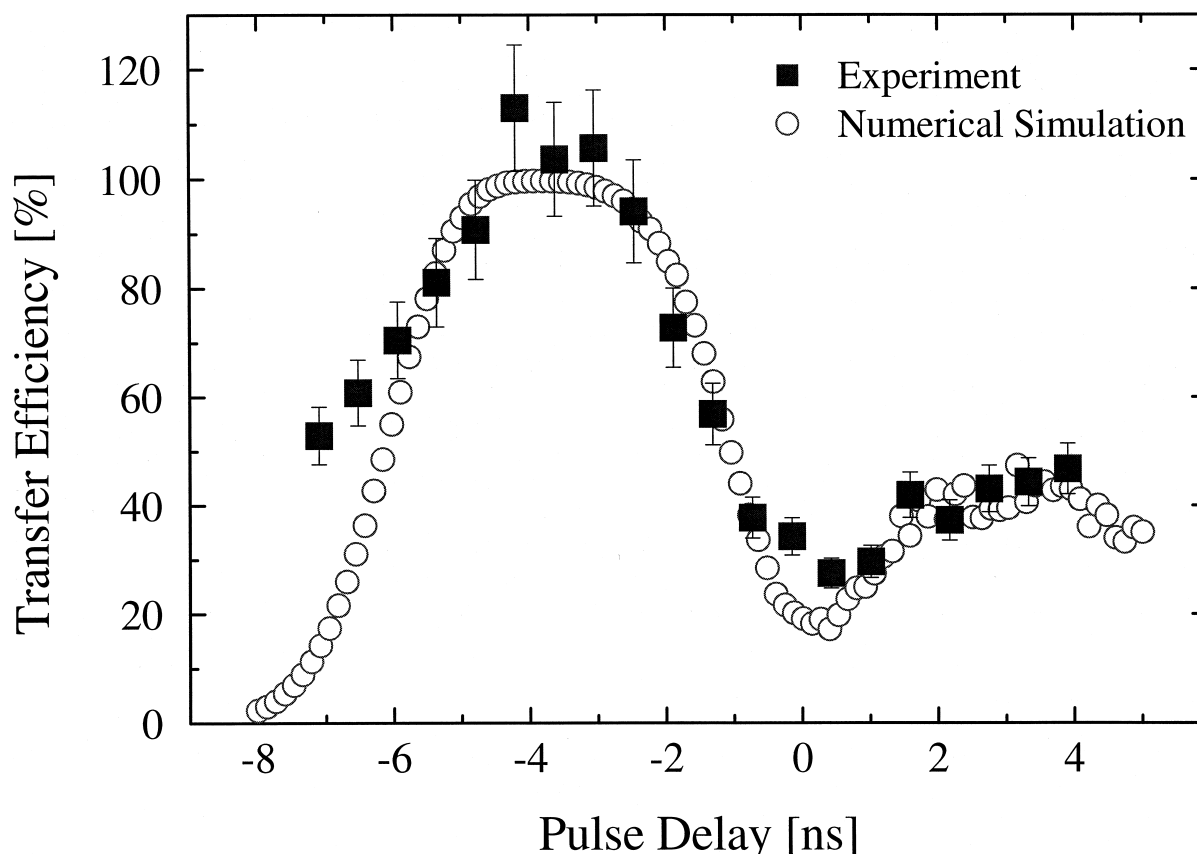
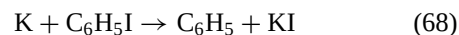


FIGURE 25 Efficiency of the transfer of population in SO_2 by STIRAP as a function of time delay between the pump and Stokes laser pulses. [Reproduced with permission from Halfmann, T., and Bergmann, K. (1996). *J. Chem. Phys.* **104**, 7068. Copyright American Institute of Physics.]

molecule. In this section we show how the dipole force can be used to manipulate the external motion of a molecule. A particle with an electric dipole moment μ and a polarizability α exposed to an electric field ε has a potential energy

$$V(\theta) = -\mu\varepsilon \cos \theta - \frac{1}{2}\varepsilon^2(\alpha_{\parallel} \cos^2 \theta + \alpha_{\perp} \sin^2 \theta), \quad (67)$$

where θ is the angle between μ and ε , and $\alpha_{\parallel}$ and $\alpha_{\perp}$ are the parallel and perpendicular components of the polarizability tensor, respectively. For a static electric field the first term dominates, and the potential energy has a minimum at $\theta = 0$. Static field strengths on the order of 10^4 – 10^5 V/cm are typically required to align or orient a molecule. (*Alignment* refers to the direction of the molecular axis without regard to which end is “up,” whereas *orientation* treats a molecule as a single-headed arrow.) For example, Wei Kong used a field strength of 58 kV/cm to orient a molecular beam of BrCN. This “brute force” method of orientation has been used to study steric effects in bimolecular reactions. For example, the differential cross section for the reaction



was measured by Loesch and coworkers for different orientations of phenyl iodide. If a focusing hexapole field is used instead of a uniform dc field, it is possible to generate a beam of symmetric top molecules in a single rotational eigenstate.

Although the brute force method has produced many beautiful results, it is limited to molecules with a permanent dipole moment and is hampered by the requirements of high voltages, and, for a multipole focusing field, a complex apparatus. An alternate method of producing a strong electric field is to use a focused laser beam. In the high frequency limit, the time average of the ac electric field gives $\langle \varepsilon(t) \rangle = 0$ and $\langle \varepsilon^2(t) \rangle = \frac{1}{2} \langle \varepsilon_0^2 \rangle$, where ε_0 is the amplitude of the field. The potential energy of the molecule in this case is

$$V(\theta) = -\frac{1}{4}\varepsilon_0^2(\Delta\alpha \cos^2 \theta + \alpha_{\perp}), \quad (69)$$

where $\Delta\alpha = \alpha_{\parallel} - \alpha_{\perp}$ is the anisotropy of the polarizability. A dimensionless quantity that is useful for scaling the alignment of different molecules is given by

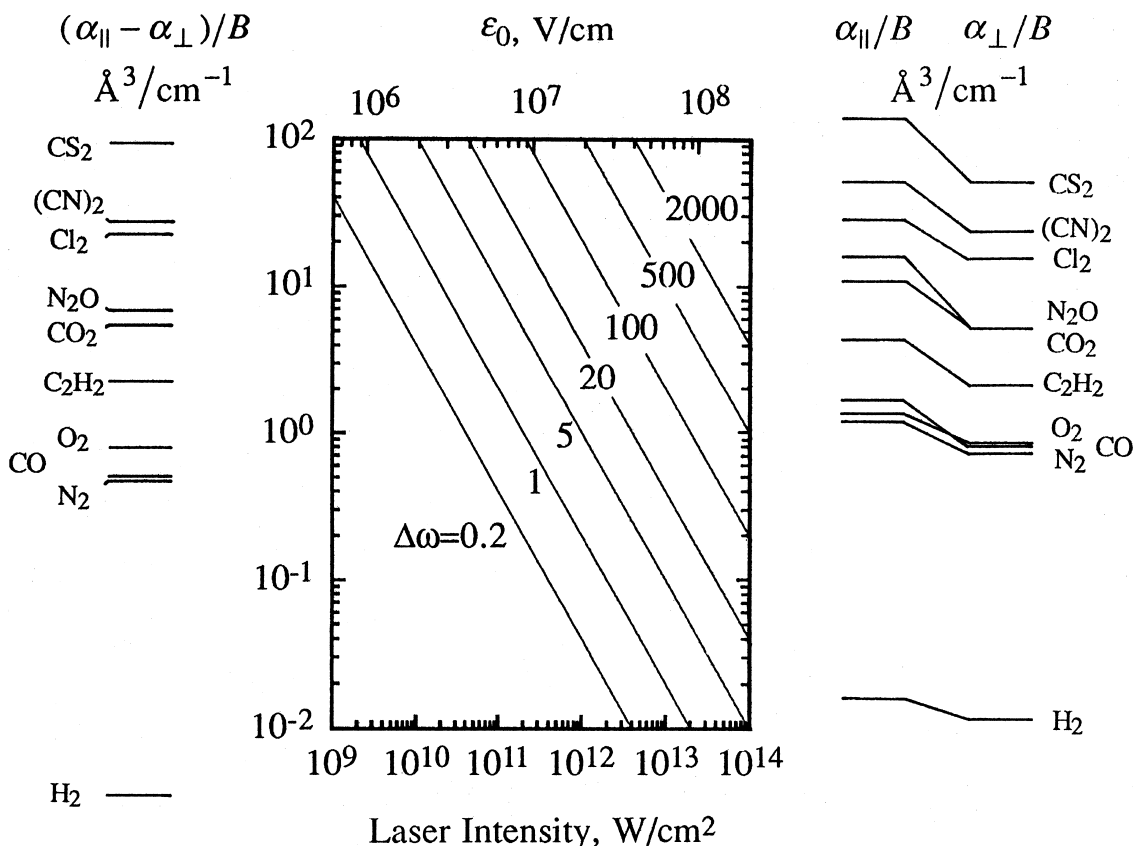


FIGURE 26 Nomogram of the dimensionless alignment parameter $\Delta\omega$ as a function of laser intensity, field strength, and polarizability anisotropy. [Reproduced with permission from Freidrich, B., and Herschbach, D. R. (1995). *J. Phys. Chem.* **99**, 15686. Copyright American Chemical Society.]

$$\omega_{\parallel(\perp)} = \alpha_{\parallel(\perp)} \epsilon_0^2 / 4B, \quad (70)$$

where B is the rotational constant. Noting that $I = 0.00265 \epsilon_0^2$, where I is the laser intensity in W/cm^2 , and ϵ_0 is in V/cm , we obtain

$$\omega_{\parallel(\perp)} = 5.28 \times 10^{-12} \alpha_{\parallel(\perp)} [\text{\AA}^3] I [\text{W/cm}^2] / B [\text{cm}^{-1}]. \quad (71)$$

The condition for strong alignment is $\Delta\omega = \omega_{\parallel} - \omega_{\perp} \cong 10$, which, for $\alpha = 10 \text{\AA}^3$ and $B = 1 \text{ cm}^{-1}$, corresponds to $I = 2 \times 10^{11} \text{ W/cm}^2$. A nomogram for $\Delta\omega$, plotted in Fig. 26, shows that many molecules may be aligned with fields readily achieved in the laboratory.

To align a molecule, it is necessary that the intensity of the aligning laser pulse lie below the threshold for multiphoton ionization. From Eq. (70) it is evident that the required intensity varies inversely with polarizability. Because molecules with low polarizabilities generally have higher ionization potentials, the conditions for laser alignment are fairly robust. An experiment demonstrating alignment utilized three laser pulses: a linearly polarized aligning pulse ($\tau = 3.5 \text{ ns}$, $I = 1.4 \times 10^{12} \text{ W/cm}^2$,

$\lambda = 1.064 \text{ }\mu\text{m}$), a circularly polarized dissociating pulse (100 fs , $3 \times 10^{12} \text{ W/cm}^2$, 688 nm), and a circularly polarized ionizing pulse (100 fs , $7 \times 10^{13} \text{ W/cm}^2$, 800 nm). The fragment ions were accelerated toward a microchannel plate, and an image produced by electrons ejected onto a phosphor screen was captured by a charge-coupled device (CCD) camera. Typical images demonstrating alignment of $\text{C}_6\text{H}_5\text{I}$ are shown in Fig. 27. The image on the left shows the anisotropic recoil of the iodine atom produced by a linearly polarized dissociation laser in the absence of an aligning pulse. The middle image shows isotropic recoil produced by a circularly polarized dissociation laser in the absence of an aligning pulse. Finally, the image on the right shows the highly anisotropic recoil produced by a circularly polarized dissociation laser in the presence of an aligning pulse. Other molecules that have been so aligned include I_2 , ICl , CS_2 , and CH_3I . If instead the aligning pulse is elliptically polarized, it is possible to align all three axes of a molecule, as was demonstrated for 3,4-dibromothiophene ($\text{C}_4\text{H}_2\text{Br}_2\text{S}$).

The average value of $\cos^2 \theta$ is a measure of the extent of alignment. Values of $\langle \cos^2 \theta \rangle > 0.9$ obtained with

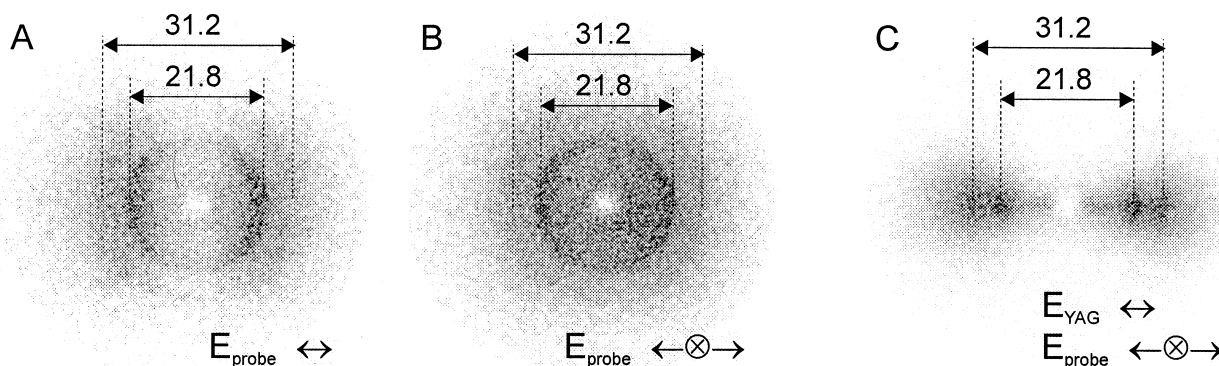


FIGURE 27 Velocity map images illustrating alignment of C_6H_5I . The images show the angular distribution of the iodine fragment for (a) a linearly polarized probe without an aligning laser, (b) a circularly polarized probe without an aligning laser, and (c) a circularly polarized probe in the presence of a linearly polarized aligning laser pulse. [Reproduced with permission from Larsen, J. J., Sakai, H., Safran, C. P., Wendt-Larsen, I., and Stapelfeldt, H. (1999). *J. Chem. Phys.* **111**, 7774. Copyright American Institute of Physics.]

the dipole force of a focused laser beam are considerably greater than what has been achieved by the brute force method. For example, pyridazine molecules, which have a permanent dipole moment of 4 Debye, when cooled to 2 K and placed in a 60 kV/cm dc field, are oriented with approximately half of the molecules restricted within a cone of 45° half-angle. In contrast, iodine molecules at the same rotational temperature and placed in a focused laser beam with an intensity of 5×10^{11} W/cm² (equivalent to 1.4×10^7 V/cm) have an induced dipole moment on the order of 1 Debye, and are aligned with half the molecules restricted to a cone of 12° and 98% of the molecules within a 45° cone.

The experiments described above used nanosecond laser pulses, which are much longer than the rotational period of the molecules. At the termination of the pulse, the pendular state that is formed relaxes adiabatically to a free-rotor eigenstate. If instead picosecond laser pulses are used, a rotational wave packet is formed by successive absorption and re-emission of photons during the laser pulse. Such wave packets are expected to display periodic recurrences of the alignment after the end of the pulse.

Laser alignment of molecules can be used to control their chemical reactions. In one example, alignment of the iodine molecule was used by Stapelfeldt and coworkers to control the spin-orbit branching ratio of its photofragments. For the I_2 molecule aligned parallel to the electric field vector of the photodissociation laser, the fragments were primarily $I(^2P_{3/2}) + I(^2P_{3/2})$, whereas for perpendicular alignment the primary products were $I(^2P_{3/2}) + I(^2P_{1/2})$. In another example studied by Corkum and coworkers, the aligned molecule was “grabbed” by the rotating polarization vector of the aligning laser and forced to move with it. Using a pair of counterrotating, circularly polarized, chirped laser pulses, the rate of rotation of a chlorine molecule was accelerated from 0 to 6 THz in 50 ps, going from near rest to angular

momentum states with $J \sim 420$. At the highest spinning rate the molecule overcame the centrifugal barrier and dissociated.

The dipole force can also be used to control the motion of the center of mass of a molecule. A focused laser beam has a radial intensity dependence $I(r)$, so that the potential energy function in Eq. (69) has a minimum at the focal point. The depth of the induced-dipole well can be orders of magnitude greater than what is commonly obtained in magneto-optic traps. For example, at an intensity of 10^{12} W/cm², the well depth for I_2 molecules is 256 K. Atoms or molecules encountering such a potential well will be deflected towards the high intensity region. As illustrated in Fig. 28, a Nd:YAG (IR) laser focused to a spot size $7 \mu\text{m}$ in diameter at an intensity of 9×10^{11} W/cm² was used to deflect a molecular beam of CS_2 . An intense (8×10^{13} W/cm²) colliding-pulse mode-locked (CPM)

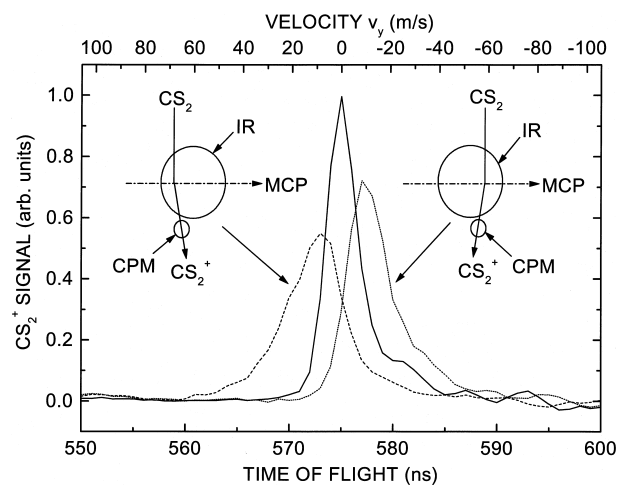


FIGURE 28 Deflection of a molecular beam of CS_2 using the dipole force of a focused laser beam. [Reproduced with permission from Stapelfeldt, H., Sakai, H., Constant, E., and Corkum, P. B. (1997). *Phys. Rev. Lett.* **79**, 2787.]

laser ionized the molecules, which were then detected by a microchannel plate (MCP). Shown in the figure are the time-of-flight spectra of the undeflected beam (solid curve) and the deflected beams (dashed and dotted curves). Molecules deflected towards the detector display an earlier arrival time than those deflected away from the detector.

VI. CONCLUDING REMARKS

Since its conception in the mid-1980s, there has been significant progress in the field of coherent control of chemical reactions. Theoretical tools for computing the optimal laser pulses for manipulating the motion of vibrational and electronic wave packets have been developed, and laboratory methods for tailoring the shapes of laser pulses are now available. These techniques have been applied to control the branching ratios of simple unimolecular ionization and dissociation reactions. Genetic algorithms have been developed and implemented for controlling more complex unimolecular reactions. The method of phase control via competing quantum mechanical paths has likewise been used to control elementary branching ratios, and a theory relating the phase lag to fundamental molecular quantities has been developed. Adiabatic passage methods have been used to achieve 100% population transfer between quantum mechanical states. Coherent methods have been developed also to control external degrees of freedom. These methods have been used to align molecules and to alter their center-of-mass translational motion.

Despite these notable successes, much remains to be done before coherent control can become a practical tool. Virtually all the successes to date have involved very simple molecules. Although learning algorithms may prove to be useful for controlling complex molecules, they have so far shed little light on the dynamics involved. Two very important problems where experiments lag far behind theory are the selective control of molecules with different chirality and the control of bimolecular reactions. A major

goal in controlling external degrees of freedom is the orientation of asymmetric molecules. In view of the major advances in laser and pulse-shaping technology over the past decade, which were barely anticipated when coherent control schemes started to emerge, it appears likely that at least some these frontier problems will be solved in the coming decade.

SEE ALSO THE FOLLOWING ARTICLES

CHEMICAL THERMODYNAMICS • DYNAMICS OF ELEMENTARY CHEMICAL REACTIONS • KINETICS (CHEMISTRY) • LASERS • NUCLEAR CHEMISTRY • PHOTOCHEMISTRY BY VUV PHOTONS • PHOTOCHEMISTRY, MOLECULAR • PROCESS CONTROL SYSTEMS • QUANTUM MECHANICS

BIBLIOGRAPHY

- Bergmann, K., Theuer, H., and Shore, B. W. (1998). "Coherent population transfer among quantum states of atoms and molecules," *Rev. Modern Phys.* **70**, 1003–1025.
- Crim, F. F. (1999). "Vibrational state control of bimolecular reactions: Discovering and directing the chemistry," *Acc. Chem. Res.* **32**, 877–884.
- Gordon, R. J., and Fujimura, Y., eds. (2001). "Advances in Multiphoton Processes and Spectroscopy, Volume 14. Quantum Control of Molecular Reaction Dynamics: Proceedings of the US–Japan Workshop," World Scientific, Singapore.
- Gordon, R. J., and Rice, S. A. (1997). "Active control of the dynamics of atoms and molecules," *Ann. Rev. Phys. Chem.* **48**, 601–641.
- Gordon, R. J., Zhu, L., and Seideman, T. (1999). "Coherent control of chemical reactions," *Acc. Chem. Res.* **32**, 1007.
- Kohler, B., Krause, J. L., Raksi, F., Wilson, K. R., Yakovlev, V. V., Whitnell, R. M., and Yan, Y. J. (1995). "Controlling the future of matter," *Acc. Chem. Res.* **28**, 133–140.
- Rice, S. A., and Zhao, M. (2000). "Optical Control of Molecular Dynamics," Wiley Interscience, New York.
- Shapiro, M., and Brumer, P., (2000). "Coherent control of atomic, molecular, and electronic processes," *Adv. Atomic Mol. Optical Phys.* **42**, 287–345.
- Tannor, D. (2002). "Introduction to Quantum Mechanics: A Time-dependent Perspective," University Science Books, Sausalito.
- Warren, S., Rabitz, H., and Dahleh, M. (1993). "Coherent control of quantum dynamics: The dream is alive," *Science* **259**, 1581–1589.



Cryogenic Process Engineering

Klaus D. Timmerhaus

University of Colorado

- I. General Applications
- II. Low-Temperature Properties
- III. Refrigeration and Liquefaction
- IV. Separation and Purification of Gases
- V. Equipment for Cryogenic Processing
- VI. Storage and Transfer Systems
- VII. Instrumentation
- VIII. Safety

GLOSSARY

Coefficient of performance Criterion used to compare refrigerator performance.

Cryobiology The use of cryogens in applications of biology.

Cryocooler Small cryogenic refrigerators and liquefiers.

Cryogen Any fluid that becomes a liquid below 125 K.

Cryogenics Term commonly associated with activities performed below 125 K.

Cryomedicine The use of cryogens in various medical and surgical applications.

Cryopumping Procedure for attaining ultrahigh vacuum by the freezing of residual gases in a chamber.

Dewar Any cryogenic container shielded by a vacuum-insulated space.

Figure of merit Criterion used to compare liquefier performance.

Helium I Normal form of helium-4 liquid existing above 2.17 K.

Helium II Superfluid form of helium-4 liquid existing below 2.17 K.

Joule–Thomson expansion Expression representing an isenthalpic throttling process.

***n*-Hydrogen** Equilibrium hydrogen existing at ambient temperatures consisting of 75% orthohydrogen and 25% parahydrogen.

MLI A high-vacuum, multilayered insulation used in cryogen storage and transfer.

Orthohydrogen Higher energy form of hydrogen resulting from the nuclear spin of the two protons in the same direction.

Parahydrogen Lower energy form of hydrogen resulting from the nuclear spin of the two protons in the opposite direction.

Regenerator Heat exchanger that periodically stores and releases heat from a matrix of high heat capacity.

Superconductivity Simultaneous disappearance of electrical resistivity and the appearance of perfect diamagnetism in a material.

Superfluid Designation for helium II existing below 2.17 K and exhibiting negligible viscosity.

CRYOGENICS is a term commonly associated with low temperatures. However, the point on the temperature scale at which refrigeration in the ordinary sense of the term ends and cryogenic engineering begins is not well defined. Most scientists and engineers working in cryogenic engineering restrict this term to a temperature range below 125 K. This is a reasonable dividing line since the normal boiling points of the more permanent gases, such as helium, hydrogen, neon, nitrogen, oxygen, and air, lie below 125 K, while the more common refrigerants have boiling points above this temperature. Thus, cryogenic process engineering is concerned with the industrial development, utilization, and improvement of low-temperature techniques, processes, and equipment.

I. GENERAL APPLICATIONS

The industrial production and utilization of temperatures below 125 K are commonly referred to as *cryogenic engineering* or *cryogenic process engineering*. This field of endeavor has grown significantly since World War II. It is now a major business in the United States with a national value in excess of \$2.5 billion annually, based on the previously defined temperature range. If the definition is broadened slightly to include the production of some petrochemicals that utilize low-temperature processing in their manufacture, such as ethylene, the annual value rapidly escalates to over \$12 billion.

An examination of cryogenic engineering shows it to be a very diverse supporting technology, a means to an end and not an end in itself. For example, oxygen, one of the most important industrial gases, is obtained by the low-temperature separation of air. Fifty percent of the oxygen produced in this manner is used by the steel industry to reduce the cost of high-grade steel, while another 20% is used in the chemical process industry to produce a variety of oxygenated compounds. Liquid hydrogen production since the mid-1950s has risen from laboratory quantities to a level of more than 250 tons/day (227,000 kg/day). Similarly, the need for liquid helium has increased by more than a factor of 15, requiring the construction of large plants to separate helium from natural gas by cryogenic means. Demands for energy have likewise accelerated the construction of tonnage base load liquefied natural gas (LNG) plants around the world and have been responsible for the associated domestic LNG industry of today with its use of peak-shaving plants.

An introduction of cryogenics would be incomplete without brief mention of some of the many current appli-

cations. For example, the phenomenon of superconductivity occurring at low temperatures has been successfully exploited in the development of high-field magnets for various uses. Space simulation is another application using a low-temperature concept. In this case, cryopumping, or the freezing of residual gases in a chamber on a cold surface, is used to provide the ultrahigh vacuum representative of outer space. This concept has been encompassed in several commercial vacuum pumps.

Freezing as a means of preserving food dates back to 1840. However, today the food industry uses large amounts of liquid nitrogen for this purpose and as a refrigerant in frozen-food transport systems. The use of cryogenics in biology and medicine has generated such interest that work in these low-temperature areas is now identified as cryobiology and cryomedicine, respectively. For example, liquid nitrogen-cooled containers are routinely used to preserve whole blood, bone marrow, and animal semen for extended periods of time. Liquid helium is used to cool the magnets in the MRI units employed by most modern hospitals. Cryogenic surgery is an acceptable procedure for curing such involuntary disorders as Parkinson's disease. Finally, one must recognize the role of cryogenics in the chemical processing industry with the treatment of natural gas streams to recover valuable heavy components or upgrade the heat content of fuel gas, the recovery of useful components from air, and the purification of various process streams.

II. LOW-TEMPERATURE PROPERTIES

Familiarity with the properties and behavior of materials used in any system operating at low temperatures is essential for proper design considerations. Since there are several significant effects among materials that become evident only at low temperatures, it is risky to obtain needed properties by an extrapolation of the variation in properties observed at ambient conditions. For example, the vanishing of specific heats, the phenomenon of superconductivity, and the onset of ductile–brittle transitions in carbon steel cannot be inferred from property measurements obtained at ambient temperatures. Accordingly, there is no substitute for test data on a truly representative sample specimen when designing for the limit of effectiveness of a cryogenic material or structure.

A. Fluid Properties

Numerous tabulations of thermodynamic property data are available in the literature. For example, a very recent tabulation of thermodynamic data by [Jacobsen, et al \(1997\)](#) covers all of the cryogenic fluids of interest. Sufficient detail on the models used for each fluid is available so

TABLE I Selected Properties of Cryogenic Liquids at Normal Boiling Point

Saturated liquid property at 0.1 MPa	Helium-4	Hydrogen ^a	Neon	Nitrogen	Air	Fluorine	Argon	Oxygen	Methane
Normal boiling point, K	4.2	20.4	27.1	77.3	78.9	85.3	87.3	90.2	111.7
Critical temperature, K	5.2	33.2	44.4	126.1	133.3	118.2	150.7	154.6	190.7
Critical pressure, MPa	0.23	1.31	2.71	3.38	3.90	5.55	4.87	5.06	4.63
Temperature at triple point, K	<i>b</i>	13.9	19.0	63.2	—	53.5	83.8	54.4	88.7
Pressure at triple point, MPa $\times 10^3$	<i>c</i>	7.2	43.0	12.8	—	0.22	68.6	0.15	10.1
Density, kg/m ³	124.9	70.9	1204	810.8	874.0	1505	1403	1134	425.0
Heat of vaporization, kJ/kg	20.7	446.3	86.6	198.4	205.1	166.5	161.6	213.1	509.7
Specific heat, kJ/kg $\cdot$ K	4.56	9.78	1.84	2.04	1.97	1.55	1.14	1.70	3.45
Viscosity, (kg/m $\cdot$ sec) $\times 10^6$	3.57	13.06	124.0	157.9	80.6	244.7	252.1	188.0	118.6
Thermal conductivity, (kJ/m $\cdot$ sec $\cdot$ K) $\times 10^3$	0.027	0.118	0.130	0.139	—	0.135	0.123	0.148	0.111
Dielectric constant	1.0492	1.226	—	1.434	—	1.43	1.52	1.4837	1.6758

^a Equilibrium hydrogen.^b λ -Point temperature, 2.17 K.^c λ -Point pressure, 5.02×10^{-3} MPa.

that the user may program the formulations in any appropriate computer language or format consistent with a particular application. Selected property data for some common cryogenics are presented in Table I. Unique properties of several of these cryogenics are noted below.

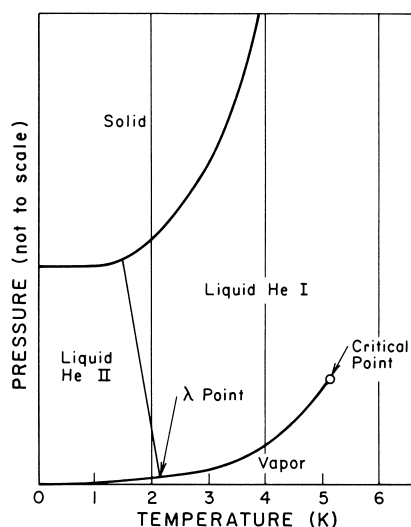
Liquid helium-4 has some very unusual properties since it can exist in two different liquid phases, namely, liquid helium I and liquid helium II (Fig. 1). The former is labeled the normal fluid, while the latter has been designated the superfluid since under certain conditions the fluid acts as if it had no viscosity. The phase transition between the two liquid phases is identified as the λ line. Intersection of the latter with the vapor-pressure curve is known as the λ point. Helium-4 has no triple point and requires a

pressure of 2.5 MPa or more even to exist as a solid below a temperature of 3 K.

Other properties of helium-4 show similar surprises. At the λ point, the specific heat of the liquid increases to a large value as the temperature is decreased through this point. Once below the λ point, the specific heat of helium II rapidly decreases to zero. The thermal conductivity of helium I, on the other hand, decreases with decreasing temperature. However, once the transition to helium II has been made, the thermal conductivity of the liquid can increase in value by as much as 10^6 that of helium I.

A unique property of hydrogen is that it can exist in two different molecular forms, namely, orthohydrogen and parahydrogen. The ortho and para forms differ in the relative orientation of the nuclear spins of the two atoms associated with the diatomic molecule. The thermodynamic equilibrium composition of the ortho and para varieties is temperature dependent. At ambient temperatures, the equilibrium mixture is 75% orthohydrogen and 25% parahydrogen and is designated as normal hydrogen. With decreasing temperatures, the thermodynamic equilibrium shifts to essentially 100% parahydrogen at 20.4 K, the normal boiling point of hydrogen. The conversion from normal hydrogen to parahydrogen is exothermic and evolves sufficient energy to vaporize $\sim 1\%$ of the stored liquid per hour. To minimize such losses in the commercial production of liquid hydrogen, a catalyst is used to effect the conversion from normal hydrogen to the thermodynamic equilibrium concentration during the liquefaction process.

The two forms of hydrogen have different specific heats. This difference, in turn, affects other thermal and transport properties of hydrogen. For example, parahydrogen gas

**FIGURE 1** Phase diagram for helium-4.

has a higher thermal conductivity than orthohydrogen gas because of the higher specific heat of the parahydrogen gas.

In contrast to other cryogenics, liquid oxygen is slightly magnetic. It is also chemically very reactive with hydrocarbon materials. It thus presents a safety problem and requires extra precautions in handling.

Fluorine is characterized by its high toxicity and extreme reactivity. The fatal concentration range for animals is $200 \text{ ppm} \times \text{hr}$, while the maximum allowable dosage for humans is usually considered to be $1 \text{ ppm} \times \text{hr}$.

B. Thermal Properties

The thermal properties of most interest at low temperatures are specific heat, thermal conductivity, and thermal expansivity.

1. Specific Heat

Specific heat can be predicted fairly accurately by mathematical models through statistical mechanics and quantum theory. For solids, the Debye model gives a satisfactory representation of the specific heat with temperature. Difficulties, however, are encountered when the Debye theory is applied to alloys and compounds. Plastics and glasses are other classes of solids that fail to follow this theory. In such cases, only experimental test data will provide sufficiently reliable specific heat values.

In general, the specific heat of cryogenic liquids decreases in a manner similar to that noted for crystalline solids as the temperature is lowered. At low pressures, the specific heat decreases with a decrease in temperature. However, at high pressures in the neighborhood of the critical, humps in the specific-heat curve are also observed for all normal cryogenics.

2. Thermal Conductivity

Adequate predictions of thermal conductivity for pure metals can be made by means of the Wiedemann–Franz law, which states that the ratio of the thermal conductivity to the product of the electrical conductivity and the absolute temperature is a constant. High-purity aluminum and copper exhibit peaks in thermal conductivity between 20 and 50 K, but these peaks are rapidly suppressed with increased impurity levels and cold work of the metal. The aluminum alloys Inconel, Monel, and stainless steel show a steady decrease in thermal conductivity with a decrease in temperature. This behavior makes these structural materials useful in any cryogenic service that requires low thermal conductivity over an extended temperature range.

All cryogenic liquids except hydrogen and helium have thermal conductivities that increase as the temperature is

decreased. For these two exceptions, the thermal conductivity decreases with a decrease in temperature. The kinetic theory of gases correctly predicts the decrease in thermal conductivity of all gases as the temperature is lowered.

3. Thermal Expansivity

The expansion coefficient of a solid can be estimated with the aid of an approximate thermodynamic equation of state for solids that equates the thermal expansion coefficient β with the quantity $\gamma C_v \rho / B$, where γ is the Grüneisen dimensionless ratio, C_v the specific heat of the solid, ρ the density of the material, and B the bulk modulus. For face-centered cubic (fcc) metals, the average value of the Grüneisen constant is ~ 2.3 . However, there is a tendency for this constant to increase with atomic number.

C. Electrical and Magnetic Properties

1. Electrical Resistivity

The electrical resistivity of most pure metallic elements at ambient and moderately low temperatures is approximately proportional to the absolute temperature. At very low temperatures, however, the resistivity (except that of superconductors) approaches a residual value almost independent of temperature. Alloys, on the other hand, have resistivities much higher than those of their constituent elements and resistance–temperature coefficients that are quite low. The electrical resistivity as a consequence is largely independent of temperature and may often be of the same magnitude as the room-temperature value.

The insulating quality of solid electrical conductors usually improves as the temperature is lowered. In fact, all the common cryogenic fluids are good electrical insulators.

2. Superconductivity

The phenomenon of superconductivity involving the simultaneous disappearance of all electrical resistance and the appearance of diamagnetism is undoubtedly the most distinguishing characteristic of cryogenics. The Bardeen–Cooper–Schrieffer (BCS) theory has been successful in accounting for most of the basic features observed of the superconducting state for low-temperature superconductors (LTS) operating below 23 K. The advent of the ceramic high-temperature superconductors (HTS), operating between 77 and 125 K, has called for modifications to existing theories that still have not been finalized. The list of materials whose superconducting properties have been measured extends into the thousands.

Three important characteristics of the superconducting state are the critical temperature, the critical magnetic

field, and the critical current. These parameters can be varied by using different materials or giving them special metallurgical treatments. For pure, unstrained metals, the normal (atmospheric) transition temperature from the superconducting state to the normal state is very sharp. For alloys, intermetallic compounds, and ceramics, the transition temperature can be quite large. Superconductivity in any of these materials, however, can be destroyed by subjecting the material either to an external or a self-induced magnetic field that exceeds a predetermined threshold field.

The alloy niobium–titanium (NbTi) and the intermetallic compound of niobium and tin (Nb_3Sn) are the most technologically advanced LTS materials presently available. Even though NbTi has a lower critical field and critical current density, it is often selected because its metallurgical properties favor convenient wire fabrication.

There are several families of high-temperature superconductors under investigation for practical magnet applications. Most of these HTS materials are copper oxide ceramics with varying oxygen contents. Because of their ceramic nature, HTS materials are quite brittle. This has introduced problems with some rather unique solutions relative to the fabrication of flexible wires that can be used in the windings of superconducting magnets.

Several of the low-temperature superconducting metals, such as lead, brass, and some solders (particularly lead–tin alloys), experience property changes when they become superconducting. Such changes can include specific heat, thermal conductivity, electrical resistance, magnetic permeability, and thermoelectric resistance. Consequently, the use of these superconducting metals in the construction of equipment for low-temperature operation must be evaluated carefully.

D. Mechanical Properties

A number of mechanical properties are of interest to the cryogenic engineer contemplating the design of a low-temperature facility. These properties include ultimate and yield strength, fatigue strength, impact strength, hardness, ductility, and elastic moduli.

1. Strength, Ductility, and Elastic Modulus

It is most convenient to classify metals by their lattice symmetry for low-temperature mechanical properties considerations. The fcc metals and their alloys are most often used in the construction of cryogenic equipment. Aluminum, copper, nickel, their alloys, and the austenitic stainless steels of the 18–8 type are fcc and do not exhibit an impact ductile-to-brittle transition at low temperatures. Generally, the mechanical properties of these metals im-

prove as the temperature is reduced. The yield strength at 20 K is considerably larger than at ambient temperature; Young's modulus is 5 to 20% larger at the lower temperatures, and fatigue properties, with the exception of 2024-T4 aluminum, are also improved at the lower temperatures. Since annealing of these metals and alloys can affect both the ultimate and yield strengths, care must be exercised under these conditions.

The body-centered cubic (bcc) metals and alloys are normally classified as undesirable for low-temperature construction. This class includes iron, the martensitic steels (low carbon and the 400 series of stainless steels), molybdenum, and niobium. If not brittle at room temperature, these materials exhibit a ductile-to-brittle transition at low temperatures. Cold working of some steels, in particular, can induce the austenite-to-martensite transition.

The hexagonal close-packed (hcp) metals exhibit mechanical properties intermediate between those of the fcc and bcc metals. For example, zinc suffers a ductile-to-brittle transition, whereas zirconium and pure titanium do not. The latter and its alloys have an hcp structure, remain reasonably ductile at low temperatures, and have been used for many applications where weight reduction and reduced heat leakage through the material have been important. However, small impurities of oxygen, nitrogen, hydrogen, and carbon can have a detrimental effect on the low-temperature ductility properties of titanium and its alloys.

Plastics increase in strength as the temperature is decreased, but this is also accompanied by a rapid decrease in elongation in a tensile test and a decrease in impact resistance. Teflon and glass-reinforced plastics retain appreciable impact resistance as the temperature is lowered. The glass-reinforced plastics also have high strength-to-weight and strength-to-thermal conductivity ratios. All elastomers, on the other hand, become brittle at low temperatures. Nevertheless, many of these materials, including rubber, Mylar, and nylon, can be used for static seal gaskets provided that they are highly compressed at room temperature before cooling.

The strength of glass under constant loading also increases with a decrease in temperature. Since failure occurs at a lower stress when the glass surface contains surface defects, the strength can be improved by tempering the surface.

III. REFRIGERATION AND LIQUEFACTION

Refrigeration in a thermodynamic process is accomplished when the process fluid absorbs heat at temperatures below that of the environment. Heat absorption at refrigerating temperatures can take place in totally different

ways. If a low-temperature liquid is formed in the process, the heat that is absorbed evaporates the liquid, and refrigeration is accomplished at constant temperature. If the refrigerator is designed to reduce the process fluid to a cold gaseous state, the heat absorbed changes the sensible heat and consequently the temperature of the fluid.

In a continuous refrigeration process, there is no accumulation of refrigerant in any part of the system. This contrasts with a gas-liquefying system, where liquid accumulates and is withdrawn. Thus, in a liquefying system, the total mass of gas that is warmed and returned to the low-pressure side of the compressor is less than the gas to be cooled by the amount liquefied, creating an unbalanced flow in the heat exchangers. In a refrigerator, the warm and cool gas flows are usually equal in the heat exchangers, except where a portion of the flow is diverted through a work-producing expander. This results in what is usually referred to as “balanced flow condition” in a heat exchanger.

A process for producing refrigeration at cryogenic temperatures usually involves equipment at ambient temperature in which the process fluid is compressed and heat is rejected to a coolant. During the ambient temperature compression process, the enthalpy and entropy are decreased. At the cryogenic temperature where heat is absorbed, the enthalpy and entropy are increased. The reduction in temperature of the process fluid is usually accomplished by heat exchange between the cooling and warming fluid followed by an expansion. This expansion may take place either through a throttling device (isenthalpic expansion), where there is only a reduction in temperature, or in a work-producing device (isentropic expansion), where both temperature and enthalpy are decreased.

A. Isenthalpic Expansion

The simple Linde cycle shown in Fig. 2a provides a good example of an isenthalpic expansion process. In this process the gaseous refrigerant is initially compressed to approximate isothermal conditions by rejecting heat to a coolant. The compressed refrigerant is cooled in a heat exchanger by the stream returning to the compressor intake until it reaches the throttling valve. Joule–Thomson cooling on expansion further reduces the temperature until a portion of the refrigerant is liquefied. For a refrigerator, the unliquefied fraction and the vapor formed by liquid evaporation from the absorbed heat Q are warmed in the heat exchanger as they are returned to the compressor intake. Figure 2b shows this process on a temperature–entropy diagram. If one applies the first law to this refrigeration cycle and assumes no heat inleaks as well as negligible kinetic and potential energy changes in the re-

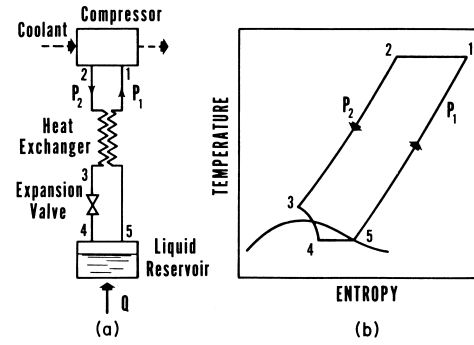


FIGURE 2 (a) Schematic for simple Linde-cycle refrigerator; (b) temperature–entropy diagram for cycle.

frigerant fluid, the refrigeration effect per unit mass of refrigerant compressed will simply be the difference in enthalpies of streams 1 and 2 of Fig. 2a. Thus, the coefficient of performance (COP) of the ideal simple Linde cycle is given by:

$$\text{COP} = \frac{Q_{\text{ref}}}{W} = \frac{h_1 - h_2}{T_1(s_1 - s_2) - (h_1 - h_2)} \quad (1)$$

where Q_{ref} is the refrigeration effect; W the work of compression; h_1 and h_2 the enthalpies at points 1 and 2, respectively; and s_1 and s_2 the entropies at points 1 and 2, respectively, of Fig. 2a.

For a simple Linde liquefier, the liquefied portion is continuously withdrawn from the reservoir, and only the unliquefied portion of the fluid is warmed in the counter-current heat exchanger and returned to the compressor. The fraction y that is liquefied is obtained by applying the first law to the heat exchanger, throttling valve, and liquid reservoir. This results in:

$$y = (h_1 - h_2)/(h_1 - h_f) \quad (2)$$

where h_f is the specific enthalpy of the liquid being withdrawn. Note maximum liquefaction occurs when the difference between h_1 and h_2 is maximized. To account for heat inleak q_L , the relation is modified to:

$$y = (h_1 - h_2 - q_L)/(h_1 - h_f) \quad (3)$$

with a resultant decrease in the fraction liquefied. The work of compression is identical to that for the simple Linde refrigerator. The figure of merit (FOM), defined as $(W/m_f)_i/(W/m_f)$, where $(W/m_f)_i$ is the work of compression per unit mass liquefied for the ideal liquefier and (W/m_f) the work of compression per unit mass liquefied for the simple Linde cycle, reduces to the expression:

$$\text{FOM} = \left[\frac{T_1(s_1 - s_f) - (h_1 - h_f)}{T_1(s_1 - s_2) - (h_1 - h_2)} \right] \left(\frac{h_1 - h_2}{h_1 - h_f} \right) \quad (4)$$

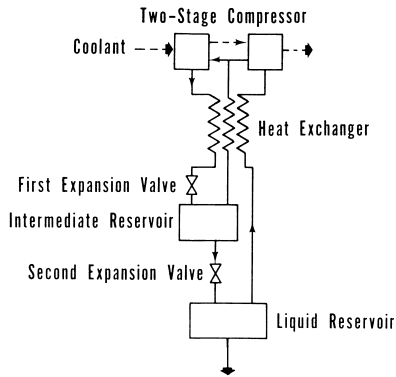


FIGURE 3 Liquefier using dual-pressure process.

Liquefaction by this cycle requires that the inversion temperature of the refrigerant be above the ambient temperature to provide cooling as the process is started. Auxiliary refrigeration is required if the simple Linde cycle is to be used to liquefy fluids whose inversion temperature is below ambient. Liquid nitrogen is the optimum auxiliary refrigerant for hydrogen and neon liquefaction systems, while liquid hydrogen is the normal auxiliary refrigerant for helium liquefaction systems.

To reduce the work of compression, a two-stage, or dual-pressure, process can be used whereby the pressure is reduced by two successive isenthalpic expansions (Fig. 3). Since the work of compression is approximately proportional to the logarithm of the pressure ratio and the Joule–Thomson cooling is roughly proportional to the pressure difference, there is a much greater reduction in compressor work than in refrigerating performance. Hence, the dual-pressure process produces a given amount of refrigeration with less energy input than the simple Linde cycle refrigerator in Fig. 2.

B. Isentropic Expansion

Refrigeration can always be produced by expanding the process fluid in an engine and causing it to do work. A schematic of a simple gas refrigerator using this principle and the corresponding temperature–entropy diagram are shown in Fig. 4. Gas compressed isothermally at ambient temperature is cooled countercurrently in a heat exchanger by the low-pressure gas being returned to the compressor intake. Further cooling takes place during the work-producing expansion. In practice, this expansion is never truly isentropic, and this is reflected by path 3–4 on the temperature–entropy diagram (Fig. 4b).

Since the temperature in a work-producing expansion is always reduced, cooling does not depend on being below the inversion temperature before expansion. In large machines, the work produced during expansion is conserved.

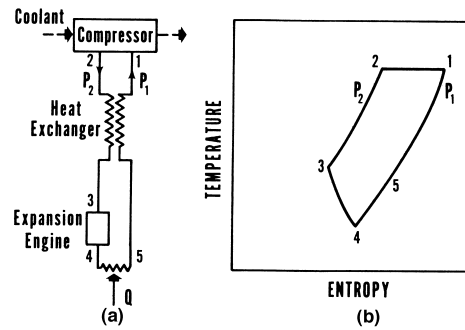


FIGURE 4 (a) Schematic for isentropic gas expansion refrigerator; (b) temperature–entropy diagram for cycle.

In small refrigerators, the energy from the expansion is usually expended in a gas or hydraulic pump or other suitable work-absorbing device.

The refrigerator in Fig. 4a produces a cold gas, which absorbs heat from 4–5 and provides a method of refrigeration for obtaining temperatures other than those at the boiling points of cryogenic fluids.

C. Combined Isenthalpic and Isentropic Expansion

It is not uncommon to combine the isentropic and isenthalpic expansions to allow the formation of liquid in the refrigerator. This is done because of the technical difficulties associated with forming liquid in the engine. The Claude cycle is an example of a combination of these methods and is shown in Fig. 5a along with the corresponding temperature–entropy diagram (Fig. 5b).

One modification of the Claude cycle that has been used extensively in high-pressure liquefaction plants for air is the Heylandt cycle. In this cycle, the first warm heat exchanger in Fig. 5a has been eliminated, permitting the inlet of the expander to operate with ambient temperature

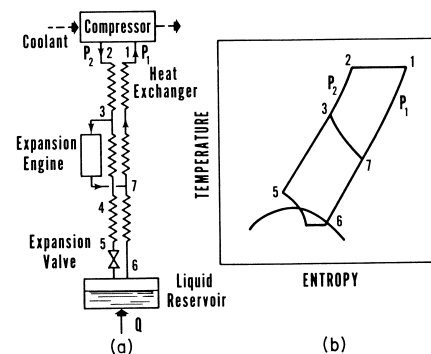


FIGURE 5 (a) Schematic for combined isenthalpic and isentropic expansion refrigerator; (b) temperature–entropy diagram for cycle.

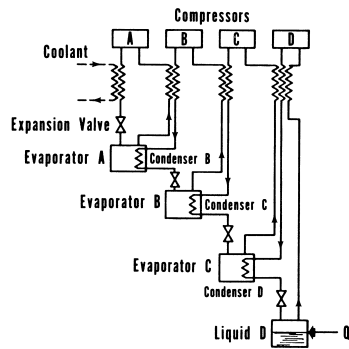


FIGURE 6 Cascade compressed vapor refrigerator.

seals, thereby minimizing lubrication problems. Another modification of the basic Claude cycle is the dual-pressure cycle utilizing the same principle as shown for the simple Linde cycle in Fig. 3. Still another extension of the Claude cycle is the Collins helium liquefier. Depending on the helium inlet pressure, from two to five expansion engines are used to provide the cooling needed in the system.

D. Mixed Refrigerant Cycle

Another cycle that has been used exclusively for large natural gas liquefaction plants is the mixed refrigerant cycle. Since this cycle resembles the classic cascade cycle in principle, it can best be understood by reference to a simplified flow sheet of that cycle presented in Fig. 6.

After purification, the natural gas stream is cooled successively by vaporization of propane, ethylene, and methane. Each of these gases, in turn, has been liquefied in a conventional refrigeration loop. Each refrigerant may be vaporized at two or three pressure levels to increase the natural gas cooling efficiency, but at a cost of considerably increased process complexity.

Cooling curves for natural gas liquefaction by the cascade process are shown in Fig. 7a,b. It is evident that the cascade cycle efficiency can be considerably improved by increasing the number of refrigerants or the number

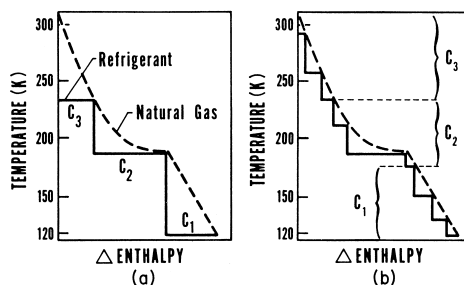


FIGURE 7 Three-level (a) and nine-level (b) cascade cycle cooling curves for natural gas.

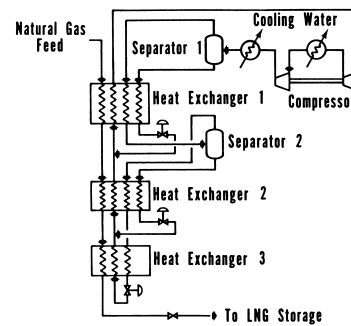


FIGURE 8 Mixed refrigerant cycle for liquefying natural gas.

of pressure levels employed. The actual work required for the nine-level cascade cycle depicted in Fig. 7b is $\sim 80\%$ of that required by the three-level cascade cycle depicted in Fig. 7a for the same throughput. The cascade system can be adapted to any cooling curve; that is, the quantity of refrigeration supplied at the various temperature levels can be chosen so that the temperature differences in the evaporators and heat exchangers approach a practical minimum (smaller temperature differences result in lower irreversibility and therefore lower power consumption).

The mixed refrigerant cycle (Fig. 8) is a variation of the cascade cycle just described and involves the circulation of a single mixed refrigerant stream, which is repeatedly condensed, vaporized, separated, and expanded. As a result, such processes require more sophisticated design methods and more complete knowledge of the thermodynamic properties of gaseous mixtures than expander or cascade cycles. Also, such processes must handle two-phase mixtures in heat exchangers. Nevertheless, simplification of the compression and heat exchange services in such cycles generally offers potential for reduced capital expenditure over conventional cascade cycles.

E. Cryocoolers

Mechanical coolers are generally classified as regenerative or recuperative. Regenerative coolers use reciprocating components that periodically move the working fluid back and forth in a regenerator. The recuperative coolers, on the other hand, use countercurrent heat exchangers to perform the heat-transfer operation. The Stirling and Gifford-McMahon cycles are typically regenerative coolers, while the Joule-Thomson and Brayton cycles are associated with recuperative coolers.

The past few years have witnessed an enhanced interest in pulse tube cryocoolers following the achievement by TRW of high-efficiency, long-life pulse tube cryocoolers based on the flexure-bearing, Stirling-cooler compressors developed at Oxford University. This interest has initiated

the development of long-life, low-cost cryocoolers for the emerging high-temperature superconductor electronic market.

During this same time period, hydrogen sorption cryocoolers have achieved their first successful operation in space, and closed-cycle, helium, Joule–Thomson cryocoolers have continued to make progress in promising long-life space applications in the 4 K temperature range. In the commercial area, Gifford–McMahon cryocoolers with rare earth regenerators have made significant progress in opening up the 4 K market.

Mixtures of highly polar gases are receiving considerable attention as refrigerants for Joule–Thomson (J–T) cycles since the magnitude of the J–T coefficient increases with nonideality of the gas. New closed-cycle J–T or throttle-cycle refrigerators have taken advantage of these mixed refrigerants to achieve low-cost cryocooler systems in the 65 to 80 K temperature range. Micro-miniature J–T cryocoolers have also been developed over the past decade using these mixed refrigerants. Fabrication of these cryocoolers uses a photolithography process in which gas channels for the heat exchangers, expansion capillary, and liquid reservoir are etched on planar glass substrates that are fused together to form the sealed refrigerator. These microminiature refrigerators have been fabricated in a wide range of sizes and capacities.

Because of the rapidly increasing availability of cryocoolers, numerous new applications have become possible; many of these involve infrared imaging systems, spectroscopy, and high-temperature superconductors in the medical and communication fields. Many of these applications have required additional control of cryocooler-generated vibration and EMI susceptibility.

F. Comparison of Refrigeration and Liquefaction Systems

A thermodynamic measure of the quality of a low-temperature refrigeration and liquefaction system is its reversibility. The second law, or more precisely the entropy increase, is an effective guide to the degree of irreversibility associated with such a system. However, to obtain a clearer picture of what these entropy increases mean, it has become convenient to relate such an analysis to the additional work required to overcome these irreversibilities. The fundamental equation for such an analysis is

$$W = W_{\text{rev}} + T_0 \sum \dot{m} \Delta s \quad (5)$$

where the total work W is the sum of the reversible work W_{rev} plus a summation of the losses in availability for vari-

ous steps in the analysis. Here, T_0 is the reference temperature (normally ambient), $\dot{m}$ the mass flow rate through the system, and Δs the change in entropy through the system.

Numerous analyses and comparisons of refrigeration and liquefaction cycles are presented in the literature. Great care must be exercised in accepting these comparisons since it is quite difficult to place all processes on a strictly comparable basis. Many assumptions are generally made in the course of these calculations, and these can have considerable effect on the conclusions. Assumptions that generally have to be made include heat leak, temperature differences in the exchangers, efficiencies of compressors and expanders, number of stages of compression, fraction of expander work recovered, state of expander exhaust, purity and condition of inlet gases, and pressure drop in the various streams. In view of this fact, differences in power requirements of 10 to 20% can be due to differences in assumed variables and can negate the advantage of one cycle over another. A comparison that demonstrates this point rather well is shown in Table II, which lists some common liquefaction systems described earlier using air as the working fluid and based on an inlet gas temperature and pressure of 294.4 K and 0.1 MPa, respectively.

IV. SEPARATION AND PURIFICATION OF GASES

The major industrial application of low-temperature processes involves the separation and purification of gases. Much of the commercial oxygen and nitrogen, and all the neon, argon, krypton, and xenon, are obtained by the distillation of liquid air. Commercial helium is separated from helium-bearing natural gas by a well-established low-temperature process. Cryogenics has also been used commercially to separate hydrogen from various sources of impure hydrogen. The low-boiling, valuable components of natural gas—namely, ethane, ethylene, propane, propylene, and others—are recovered and purified by various low-temperature schemes.

The separation of these gases is dictated by the Gibbs phase rule. The degree to which they separate is based on the physical behavior of the liquid and vapor phases. This behavior is governed, as at ambient temperatures, by Raoult's and Dalton's laws.

A. Air Separation

The simplest air separation device is the Linde single-column system, which utilizes the simple Linde

TABLE II Comparison of Several Liquefaction Systems Using Air as the Working Fluid^a

Air liquefaction system ^b	Liquid yield ($y = \dot{m}_l/\dot{m}$)	Work per unit mass liquefied, (kJ/kg)	Figure of merit
Ideal reversible system	1.000	715	1.000
Simple Linde system, $p_2 = 20$ MPa, $\eta_c = 100\%$, $\varepsilon = 1.0$	0.086	5240	0.137
Simple Linde system, $p_2 = 20$ MPa, $\eta_c = 70\%$, $\varepsilon = 0.95$	0.061	10620	0.068
Simple Linde system observed	—	10320	0.070
Precooled simple Linde system, $p_2 = 20$ MPa, $T_3 = 228$ K, $\eta_c = 100\%$, $\varepsilon = 1.00$	0.179	2240	0.320
Precooled simple Linde system, $p_2 = 20$ MPa, $T_3 = 228$ K, $\eta_c = 70\%$, $\varepsilon = 0.95$	0.158	3700	0.194
Precooled simple Linde system, observed	—	5580	0.129
Linde dual-pressure system, $p_3 = 20$ MPa, $p_2 = 6$ MPa, $i = 0.8$, $\eta_c = 100\%$, $\varepsilon = 1.00$	0.060	2745	0.261
Linde dual-pressure system, $p_3 = 20$ MPa, $p_2 = 6$ MPa, $i = 0.8$, $\eta_c = 70\%$, $\varepsilon = 0.95$	0.032	8000	0.090
Linde dual-pressure system, observed	—	6340	0.113
Linde dual-pressure system, precooled to 228 K, observed	—	3580	0.201
Claude system, $p_2 = 4$ MPa, $x = \dot{m}_e/\dot{m} = 0.7$, $\eta_c = \eta_e = 100\%$, $\varepsilon = 1.00$	0.260	890	0.808
Claude system, $p_2 = 4$ MPa, $x = \dot{m}_e/\dot{m} = 0.7$, $\eta_c = 70\%$, $\eta_{e,ad} = 80\%$, $\eta_{e,m} = 90\%$, $\varepsilon = 0.95$	0.189	2020	0.356
Claude system, observed	—	3580	0.201
Cascade system, observed	—	3255	0.221

^a Inlet conditions of 294.4 K and 0.1 MPa.^b η_c denotes compressor overall efficiency; η_e expander overall efficiency; $\eta_{e,ad}$ expander adiabatic efficiency; $\eta_{e,m}$ expander mechanical efficiency; and ε heat exchanger effectiveness. $i = \dot{m}_1/\dot{m}$ is mass in intermediate stream divided by mass through compressor, and $x = \dot{m}_e/\dot{m}$ is mass through expander divided by mass through compressor.

liquefaction cycle considered earlier but with a rectification column substituted for the liquid reservoir. (Since it is immaterial how the liquid is to be furnished to the column, any of the other liquefaction cycles could have been used in place of the simple Linde cycle.)

Although the oxygen product purity is high from a simple single-column separation scheme, the nitrogen effluent stream always contains about 6 to 7% oxygen. In other words, approximately one-third of the oxygen liquefied as feed to the column is lost in the nitrogen stream. This inherent loss of a valuable product in the single-column operation is not only undesirable but highly wasteful in terms of compression requirements.

This problem was solved by the introduction of the Linde double-column system. Two rectification columns are placed one on top of the other (hence the name double-column system). In this system, liquid air is introduced at an intermediate point in the lower column. A condenser–evaporator at the top of the lower column provides the reflux needed for the rectification process to obtain essentially pure nitrogen at this point. In order for the column to also deliver pure oxygen, the oxygen-rich liquid (~45% oxygen), from the boiler in the lower column is introduced at an intermediate level in the upper column. The reflux and the rectification process in the upper column produce pure oxygen at the bottom and

pure nitrogen at the top (provided that argon and the rare gases have been previously removed). More than enough liquid nitrogen is produced in the lower column, so that some may be withdrawn and introduced in the upper column as needed reflux. Since the condenser must condense nitrogen vapor in the lower column by evaporating liquid oxygen in the upper column, it is necessary to operate the lower column at a higher pressure, ~ 0.5 MPa, while the upper column is operated at ~ 0.1 MPa. This requires throttling to reduce the pressure of the fluids from the lower column as they are transferred to the upper column.

The processes used in industrial air-separation plants have changed very little in basic principle during the past 25 years. After cooling the compressed air to its dew point in a main heat exchanger by flowing counter current to the products of separation, the air feed, at an absolute pressure of about 6 MPa, is separated in a double distillation column. This unit is kept cold by refrigeration developed in a turbine, which expands a flow equivalent to between 8 and 15% of the air-feed stream down to approximately atmospheric pressure.

Figure 9 shows a modern air-separation plant with front-end cleanup and product liquefaction. Production of such plants can exceed 2800 tons per day of liquid oxygen with an overall efficiency of about 15 to 20% of the theoretical optimum. The recent introduction of molecular sieve technology has provided an arrangement that increases the product to about 85% of the air input to the compressor. Thus, there has been a strong tendency over the past decade to retrofit older air-separation plants with this new arrangement to improve the process.

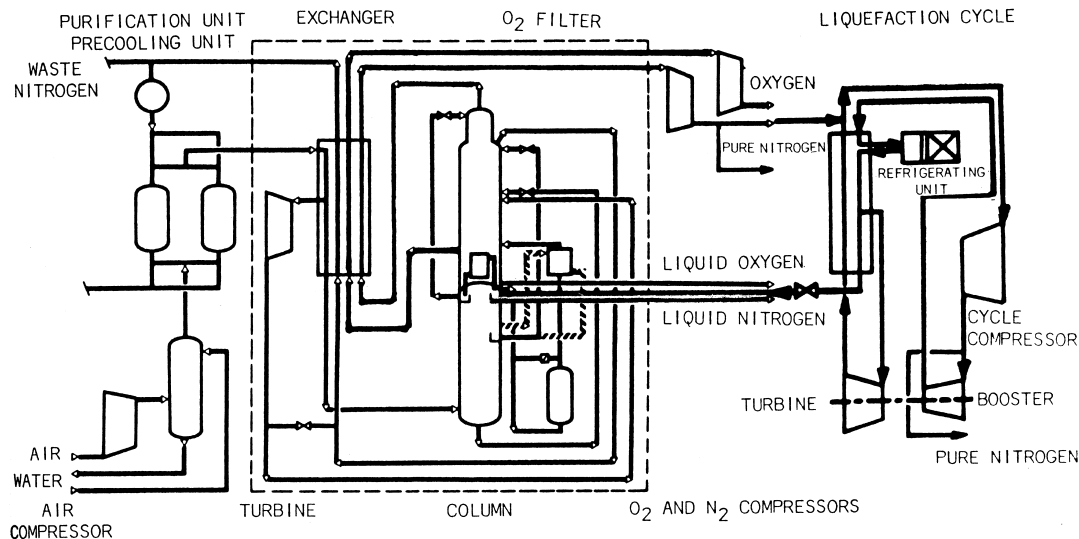


FIGURE 9 Schematic of a modern air-separation plant.

B. Rare Gas Recovery

Argon, neon, krypton, and xenon are recovered as products in commercial air separation plants. Since atmospheric air contains 0.93% argon with a boiling point intermediate between those of nitrogen and oxygen, the argon will appear as an impurity in either or both the nitrogen and the oxygen product of an air separation plant. Thus, removal of the argon is necessary if pure oxygen and nitrogen are desired from the air separation.

Figure 10 illustrates the scheme for removing and concentrating the argon. The upper column is tapped at the level where the argon concentration is highest in the column. Gas rich in argon is fed to an auxiliary column, where the argon is separated, and the remaining oxygen and nitrogen mixture is returned to the appropriate level in the primary column. The yield for this type of plant is about 50% of the atmospheric argon. The crude argon product generally contains 45% argon, 50% oxygen, and 5% nitrogen. The oxygen is readily removed by chemical reduction or adsorption. The remaining nitrogen impurity is of no consequence if the argon is to be used for filling incandescent lamps. However, for shielded-arc welding, the nitrogen must be removed by another rectification column.

Since helium and neon have boiling points considerably below that of nitrogen, these gases will collect on the nitrogen side of the condenser-reboiler associated with the double-column air separation system. Recovery of these gases is accomplished by periodic venting of a small portion of the gas from the dome of the condenser and transfer to a small condenser-rectifier refrigerated with

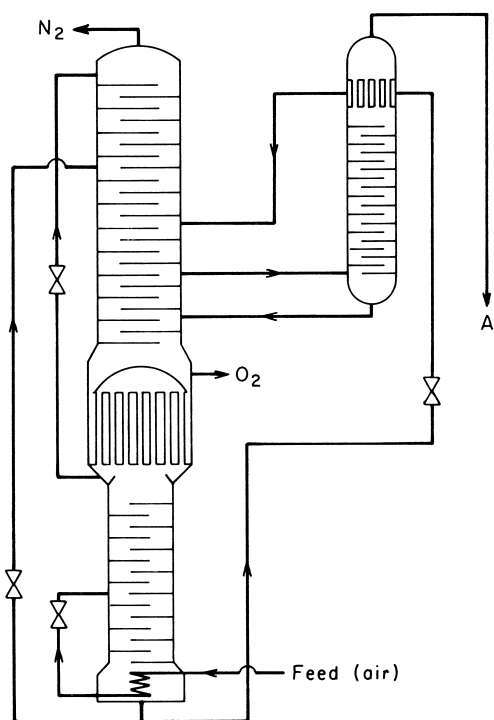


FIGURE 10 Air-separation plant with auxiliary argon separation column.

liquid nitrogen from the condenser of the double column. The resulting crude helium and neon are further purified by a series of charcoal adsorption units to provide high-purity neon.

The concentrations of krypton and xenon in atmospheric air are quite small. Thus, a very large amount of air has to be processed to produce an appreciable amount of these rare gases. Since the boiling points of krypton and xenon are higher than those of oxygen, these two components in atmospheric air tend to collect in the oxygen product of the double column. To recover these rare gases, liquid oxygen from the reboiler of the upper column is first sent to an auxiliary condenser-boiler to increase the concentration of the krypton and xenon. The product is further concentrated in another auxiliary rectification column before being vaporized and passed through a catalytic furnace to combine any remaining hydrocarbons with oxygen. The resulting water vapor and carbon dioxide are removed by a caustic trap, and the krypton and xenon are adsorbed in a silica gel trap. The krypton and xenon are then separated either in a small rectification column or by a series of adsorptions and desorptions on activated charcoal.

C. Helium Recovery

Most of the helium produced in the United States is obtained by recovering this component from helium-rich

natural gas. Fortunately, since the major constituents of natural gas have boiling points considerably higher than that of helium, the separation can be accomplished with condenser–evaporators rather than with the more expensive rectification columns.

A typical scheme for separating helium from natural gas was pioneered by the U.S. Bureau of Mines. In this scheme, the natural gas is compressed to 4.13 MPa and treated to remove water vapor, carbon dioxide, and hydrogen sulfide. The purified stream is then partially condensed by the returning low-pressure, cold natural gas stream, throttled to a pressure of 1.72 MPa, and further cooled with cold nitrogen vapor in a heat exchanger–separator, where 98% of the gas is liquefied. The cold nitrogen vapor, supplied by an auxiliary refrigeration system, not only provides the necessary cooling but also causes some rectification of the gas phase in the heat exchanger, thereby increasing the helium content. The remaining vapor phase, consisting of about 60% helium and 40% nitrogen with a very small amount of methane, is warmed to ambient temperatures and sent to temporary storage pending further purification. The liquid phase, having been depleted of helium, is used to furnish the refrigeration required to cool and condense the incoming high-pressure gas. The process is completed by recompressing the stripped natural gas and returning it to the natural gas pipeline.

Purification of the crude helium is accomplished by compressing the gas to 18.6 MPa and cooling it first in a heat exchanger and then in a separator that is immersed in a bath of liquid nitrogen. In the separator nearly all of the nitrogen in the crude helium is condensed and removed as liquid. This liquid contains some dissolved helium, which is largely removed and returned to the process gas by reducing the pressure to 1.7 MPa and separating the resultant liquid and vapor phases in a nitrogen maker. Helium from the separator has a purity of ~98.5%. The final purification is accomplished by passing the cold helium through charcoal adsorption purifiers to remove the remaining nitrogen.

D. Natural Gas Processing

The need to recover increasing amounts of valuable hydrocarbon feedstocks from natural gas streams has resulted in expanded use of low-temperature processing of these streams. The majority of such natural gas processing is now accomplished using a turboexpander in a modified isentropic expansion cycle with feed gas normally available from 1 to 10 MPa. The first step is to dry the gas to dew points of 200 K and lower. After drying, the feed gas is cooled with cold residue gas. Liquid produced at this point is separated before entering the expander and sent to the condensate stabilizer. The gas from the separator

flows to the expander. The expander exhaust stream can contain as much as 20 wt% liquid. This two-phase mixture is sent to the top section of the stabilizer, which separates the two phases. The liquid is used as reflux in this unit, while the cold gas exchanges heat with fresh feed and is recompressed by the expander-driven compressor. Many variations of this cycle are possible and have been used in actual plants.

E. Purification Schemes

The type and amount of impurities to be removed from a gas stream depend entirely on the type of process involved. For example, in the production of tonnage oxygen, various impurities must be removed to avoid plugging of the cold-process lines or to avoid buildup of hazardous contaminants. The impurities in air that contribute most to plugging are water and carbon dioxide. Helium, hydrogen, and neon, on the other hand, accumulate on the condensing side of the oxygen reboiler and reduce the rate of heat transfer unless removed by intermittent purging. The buildup of acetylene, on the other hand, can prove to be dangerous even if its feed concentration in the air does not exceed 0.04 ppm.

Refrigeration purification is a relatively simple method for removing water, carbon dioxide, and certain other contaminants from a process stream by condensation or freezing. (Either regenerators or reversing heat exchangers can be used for this purpose since a flow reversal is periodically necessary to reevaporate and remove the solid deposits.) The effectiveness of this method depends on the vapor pressure of the impurities relative to that of the major components of the process stream at the refrigeration temperature. Thus, assuming ideal gas behavior, the maximum impurity content in a gas stream after refrigeration would be inversely proportional to its vapor pressure. However, due to the departure from ideality at higher pressures, the impurity level generally will be considerably higher than that predicted for the ideal situation. Familiarity with these deviations is necessary if problems are to be avoided with this purification method.

One of the most common low-temperature methods for removing impurities involves the use of selective solid adsorbents. Such materials as silica gel, carbon, and synthetic zeolites (molecular sieves) are widely used as adsorbents because of their extremely large effective surface areas. Most of the gels and carbons have pores of various sizes in a given sample, but the synthetic zeolites can be manufactured with closely controlled pore size openings ranging from 0.4 to 1.3 nm. This additional selectivity is useful because it permits separation of gases on the basis of molecular size.

The design of low-temperature adsorbents requires knowledge of the rate of adsorption and the equilibrium conditions that exist between the solid and the gas as a function of temperature. The data for the latter are generally available from the suppliers of such adsorbents. The rate of adsorption is generally very rapid, and the adsorption is essentially complete in a relatively narrow zone of the adsorber. In usual plant operation at least two adsorption purifiers are employed—one in service while the other is being desorbed of impurities. In some cases there is an advantage in using three purifiers—one adsorbing, one desorbing, and one being cooled, with the latter two units being in series. The cooling of the purifier is generally performed using some of the purified gas to avoid adsorption during this period.

Low-temperature adsorption systems continue to find an increasing number of applications. For example, systems are used to remove the last traces of carbon dioxide and hydrocarbons in many air-separation plants. Adsorbents are also used in hydrogen liquefaction to remove oxygen, nitrogen, methane, and other trace impurities. They are also used in the purification of helium suitable for liquefaction (grade A) and for ultrapure helium (grade AAA, 99.999% purity). Adsorption at 35 K will, in fact, yield a helium with less than 2 ppb of neon, which is the only detectible impurity in helium after this treatment.

Even though most chemical purification methods are not carried out at low temperatures, they are useful in several cryogenic gas separation systems. Ordinarily water vapor is removed by refrigeration and adsorption methods. However, for small-scale purification, the gas can be passed over a desiccant, which removes the water vapor as water of crystallization. In the krypton–xenon purification system, carbon dioxide is removed by passage of the gas through a caustic, such as sodium hydroxide, to form sodium carbonate.

When oxygen is an impurity, it can be removed by reaction of the oxygen in the presence of a catalyst with hydrogen to form water. The latter then is removed by refrigeration or adsorption. Palladium and metallic nickel have proved to be effective catalysts for the hydrogen–oxygen reaction.

V. EQUIPMENT FOR CRYOGENIC PROCESSING

The achievement and utilization of low temperatures require the use of various specialized pieces of equipment including compressors, expanders, heat exchangers, pumps, transfer lines, and storage tanks. As a general rule,

design principles applicable at ambient temperatures are also valid for low-temperature design. However, underlying each aspect of such a design must be a thorough understanding of the effect of temperature on the properties of the fluids being handled and the materials of construction being selected.

A. Compression Systems

Compression power accounts for more than 80% of the total energy required in the production of industrial gases and the liquefaction of natural gas. In order to minimize the cost and maintenance of cryogenic facilities, special care must be exercised to select the appropriate compression system. The three major types of compressors widely used today are reciprocating, centrifugal, and screw. Currently, there is no particular type of compressor that is generally preferred for all applications. The final selection will ultimately depend on the specific application as well as the effect of plant site and existing facilities.

1. Reciprocating Compressors

The key feature of reciprocating compressors is their adaptability to a wide range of volumes and pressures with high efficiency. Some of the largest units for cryogenic gas production range up to 15,000 bhp (1.12×10^4 kW) and use the balanced-opposed machine concept in multistage designs with synchronous motor drive. When designed for multistage, multiservice operation, these units incorporate manual or automatic, fixed or variable, volume clearance pockets and externally actuated unloading devices where required. Balanced-opposed units not only minimize vibrations, resulting in smaller foundations, but also allow compact installation of coolers and piping, further increasing the savings.

Operating speeds of larger units are as high as 277 rpm with piston speeds for air service up to 4.3 m/sec. The larger compressors with provision for multiple services reduce the number of motors or drivers and minimize the accessory equipment, resulting in lower maintenance cost.

Compressors for oxygen service are characteristically operated at lower piston speeds, of the order of 3.3 m/sec. Maintenance of these machines requires rigid control of cleaning procedures and inspection of parts to ensure the absence of oil in the working cylinder and valve assemblies.

Engine drivers of the variable-speed type can generally operate over a 100 to 50% variation in the design speed with little loss in operating efficiency since compressor fluid friction losses decrease at the lower revolutions per minute.

2. Centrifugal Compressors

Technological advances achieved in centrifugal compressor design have resulted in improved high-speed compression equipment with capacities exceeding 280 m³/sec in a single unit. As a consequence of their high efficiency, better reliability, and design upgrading, centrifugal compressors have become accepted for low-pressure cryogenic processes such as air-separation and base-load LNG plants.

Separately driven centrifugal compressors are adaptable to low-pressure cryogenic systems because they can be coupled directly to steam turbine drives, are less critical from the standpoint of foundation design criteria, and lend themselves to gas turbine or combined cycle applications. Isentropic efficiencies of 80 to 85% are usually obtained.

3. Screw Compressors

Most screw compressors are of the oil-lubricated type. There are two types—the semihermetic and the open-drive type. In the former, the motor is located in the same housing as the compressor, while in the latter the motor is located outside of the compressor housing and thus requires a shaft seal. The only moving parts in screw compressors are two intermeshing helical rotors. The rotors consist of one male lobe, which functions as a rolling piston, and a female flute, which acts as a cylinder. Since rotary screw compression is a continuous positive-displacement process, no surges are created in the system.

Screw compressors require very little maintenance because the rotors turn at conservative speeds and are well lubricated with coolant oil. Fortunately, most of the oil can easily be separated from the gas in screw compressors. Typically, only small levels of impurities of between 1.0 and 2.0 ppm by weight remain in the gas after compression. Charcoal filters can be used to reduce the impurities below this level.

A major advantage of screw compressors is that they permit the attainment of high-pressure ratios in a single mode. To handle these same large volumes with a reciprocating compressor would require a double-stage unit. Because of this and other advantages, screw compressors are now preferred over reciprocating compressors for helium refrigeration and liquefaction applications. They are competitive with centrifugal compressors in other applications as well.

B. Expansion Devices

The primary function of a cryogenic expansion device is to reduce the temperature of the gas to provide useful

refrigeration for the process. In expansion engines, this is accomplished by converting part of the energy of the high-pressure gas stream into mechanical work. This work in large cryogenic facilities is recovered and utilized to reduce the overall compression requirements of the process. On the other hand, cooling of a gas can also be achieved by expanding the gas through an expansion valve (provided that its initial temperature is below the inversion temperature of the gas). The cooling here is accomplished by converting part of the energy of the high-pressure gas stream into kinetic energy. No mechanical work is obtained from such an expansion.

Expanders are of either the reciprocating or the centrifugal type. With the rapid growth of tonnage in cryogenic processes, centrifugal expanders have gradually displaced the reciprocating type in large plants. However, the reciprocating expander is still popular for those processes where the inlet temperature is very low, such as for hydrogen or helium gas. Units up to 3600 hp (2685 kW) have been put in service for nitrogen expansion in liquid hydrogen plants, while nonlubricated expanders with exhausts well below 33.3 K are being used in liquid hydrogen plants developed for the space program.

1. Reciprocating Expanders

Generally, reciprocating expanders are selected when the inlet pressure and pressure ratio are high and when the volume of gas handled is low. The inlet pressure to expansion engines used in air-separation plants varies between 4 and 20 MPa, while capacities range from 0.1 to 3 m³/sec. Isentropic efficiencies achieved are from 70 to 80%.

The design features of reciprocating expanders employed in low-temperature processes include rigid, guided cam-actuated valve gears, renewable hardened valve seats, helical steel or air-springs, and special valve packings that eliminate leakage. Cylinders are normally steel forgings effectively insulated from the rest of the structure. Removable cylinder liners of Micarta or similar nonmetallic material and floating piston design offer wear resistance and good alignment in operation. Piston rider rings serve as guides for the piston. Nonmetallic rings are used for nonlubricated service. Both horizontal and vertical design, and one and two cylinder versions, have been used successfully.

Reciprocating expanders, in normal operation, should not accept liquid in any form during the expansion cycle. However, the reciprocating device can tolerate some liquid for short periods of time provided that none of the constituents freeze out in the expander cylinder and cause serious mechanical problems. If selected design conditions indicate possibilities of entering the liquid and especially the triple point range on expansion during normal operation, then inlet pressure and temper-

ature must be revised or thermal efficiency modified accordingly.

C. Centrifugal Expanders

Turboexpanders are classified as either axial or radial. Most turboexpanders built today are of the radial type because of their generally lower cost and reduced stresses for a given tip speed. This permits them to run at higher speeds with higher efficiencies and lower operating costs. On the other hand, axial flow expanders are more suitable for multistage expanders since they permit a much easier flow path from one stage to the next. Where low flow rates and high enthalpy reductions are required, an axial flow two-stage expander is generally used with nozzle valves controlling the flow. For example, in the processing of ethylene, gas leaving the demethanizer is normally saturated, and expansion conditions result in a liquid product coming out of the expander. Since up to 15 to 20% liquid at the isentropic end point can be handled in actual flow impulse turbine expanders, recovery of ethylene is feasible by the procedure. Depending on the initial temperature and pressure into the expander and the final exit pressure, good flow expanders are capable of reducing the enthalpy of an expanded fluid by between 175 and 350 kJ/kg, and this may be multistaged. The change in enthalpy drop can be automatically regulated by turbine speed. The development of highly reliable and efficient turboexpanders has made today's large-tonnage air-separation plants and base-load LNG facilities a reality. Notable advances in turboexpander design for these applications center on improved bearings, lubrication, and wheel and rotor design to permit nearly ideal rotor assembly speeds with good reliability. Pressurized labyrinth sealing systems use dry seal gas under pressure mixed with cold gas from the process to provide seal output temperatures above the frost point. Seal systems for oxygen compressors are more complex than for air or nitrogen and must prevent oil carryover to the processed gas. By the combination of variable-area nozzle grouping or partial admission of multiple nozzle grouping, efficiencies up to 85% have been obtained with radial turboexpanders.

Turboalternators were developed in the 1960s to improve the efficiency of small cryogenic refrigeration systems. This is accomplished by converting the kinetic energy in the expanding fluid to electrical energy, which in turn is transferred outside of the system, where it can be converted to heat and dissipated to an ambient heat sink.

D. Expansion Valve

The expansion valve or Joule-Thomson valve, as it is often called, is an important component in any liquefaction

system, although not as critical a component as the others mentioned in this section. In simplest terms the expansion valve resembles a normal valve that has been modified to handle the flow of cryogenic fluids. These modifications include exposing the high-pressure stream to the lower part of the valve seat to reduce sealing problems, a valve stem that has been lengthened and constructed of a thin-walled tube to reduce heat transfer, and a stem seal that is accomplished at ambient temperatures.

E. Heat Exchangers and Regenerators

One of the more critical components of any low-temperature liquefaction and refrigeration system is the heat exchanger. This point is readily demonstrated by considering the effect of the heat exchanger effectiveness on the liquid yield of nitrogen in a simple Linde-cycle liquefaction process operating between a lower and upper pressure of 0.1 and 10 MPa, respectively. The liquid yield under these conditions will be zero whenever the effectiveness of the heat exchanger falls below 90%. (Heat exchanger effectiveness is defined as the ratio of the actual heat transfer to the maximum possible heat transfer.)

Fortunately, most cryogens, with the exception of helium II, behave as “classical” fluids. As a result, it has been possible to predict their behavior by using well-established principles of mechanics and thermodynamics applicable to many room-temperature fluids. In addition, this has permitted the formulation of convective heat transfer correlations for low-temperature designs of simple heat exchangers that are similar to those used at ambient conditions and utilize such well-known dimensionless quantities as the Nusselt, Reynolds, Prandtl, and Grashof numbers.

However, the requirements imposed by the need to operate more efficiently at low temperatures has made the use of simple exchangers impractical in many cryogenic applications. In fact, some of the important advances in cryogenic technology are directly related to the development of rather complex but very efficient types of heat exchangers. Some of the criteria that have guided the development of these units for low-temperature service are (1) a small temperature difference at the cold end of the exchanger to enhance efficiency, (2) a large ratio of heat-exchange surface area to heat-exchanger volume to minimize heat leak, (3) a high heat-transfer rate to reduce surface area, (4) a low mass to minimize start-up time, (5) multichannel capability to minimize the number of exchangers, (6) high-pressure capability to provide design flexibility, (7) a low or reasonable pressure drop to minimize compression requirements, and (8) minimum maintenance to minimize shutdowns.

The selection of an exchanger for low-temperature operation is normally determined by process design require-

ments, mechanical design limitations, and economic considerations. The principal industrial exchangers finding use in cryogenic applications are coiled-tube, plate-fin, reversing, and regenerator types.

1. Coiled-Tube Exchangers

Construction of these widely used heat exchangers involves winding a large number of tubes in helix fashion around a central core mandrel with each exchanger containing many layers of tubes, both along the principal and radial axes. Pressure drops in the coiled tubes are equalized for each specific stream by using tubes of equal length and carefully varying the spacing of these tubes in the different layers. A shell is fitted over the outermost tube layer, and this shell together with the outside surface of the core mandrel form the annular space in which the tubes are nested. Coiled-tube heat exchangers offer unique advantages, especially for those low-temperature design conditions where simultaneous heat transfer between more than two streams is desired, a large number of heat transfer units is required, and high operating pressures in various streams are encountered. The geometry of these exchangers can be varied widely to obtain optimum flow conditions for all streams and still meet heat transfer and pressure drop requirements.

Optimization of the coiled-tube heat exchanger is quite complex. There are numerous variables, such as tube and shell flow velocities, tube diameter, tube pitch, and layer spacing. Other considerations include single-phase and two-phase flow, condensation on either the tube or shell side, and boiling or evaporation on either the tube or shell side. Additional complications come into play when multicomponent streams are present, as in natural gas liquefaction, since mass transfer accompanies the heat transfer in the two-phase region.

Many empirical relationships have been developed to aid the optimization of coiled-tube exchangers under ambient conditions. Many of the same relationships are currently being used in low-temperature applications as well. A number of these relationships are tabulated in readily available cryogenic texts. However, no claim is made that these relationships will be more suitable than others for a specific design. This can be verified only by experimental measurements on the heat exchanger.

2. Plate-Fin Exchangers

These types of heat exchangers normally consist of heat-exchange surfaces obtained by stacking alternate layers of corrugated, high-uniformity, die-formed aluminum sheets (fins) between flat aluminum separator plates to form individual flow passages. Each layer is closed at the edge with solid aluminum bars of appropriate shape and size.

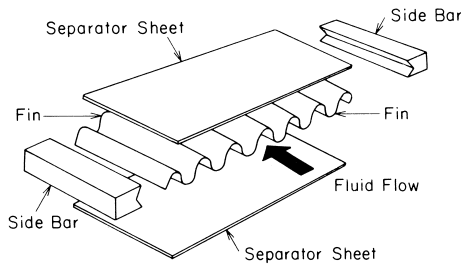


FIGURE 11 Exploded view of one layer of a plate–fin exchanger before brazing.

Figure 11 illustrates by exploded view the elements of one layer, in relative position, before being joined by a brazing operation to yield an integral rigid structure with a series of fluid flow passages. The latter normally have integral welded headers. Several sections can be connected to form one large exchanger. The main advantages of this type of exchanger are that it is compact (about nine times as much surface area per unit volume as conventional shell and tube exchangers), yet permits wide design flexibility, involves minimum weight, and allows design pressures to 6 MPa from 4.2 to 340 K.

The fins for these heat exchangers can be manufactured in a variety of configurations that can significantly alter the heat transfer and pressure drop characteristics of the exchanger. Various flow patterns can be developed to provide multipass or multistream arrangements by incorporating suitable internal seals, distributors, and external headers. The type of headers used depends on the operating pressures, the number of separate streams involved, and, in the case of a counterflow exchanger, whether reversing duty is required.

Design of the plate–fin exchanger involves selecting a geometry and surface arrangement that will give a product UA of the correct magnitude to satisfy the relation:

$$Q = UA\Delta T_e \quad (6)$$

where Q is the heat transfer rate, U the overall heat-transfer coefficient, A the area of heat transfer, and ΔT_e the equivalent temperature difference between the two streams exchanging energy.

Plate–fin exchangers can be supplied as single units or as manifolded assemblies that consist of multiple units connected in parallel or in series. Sizes of single units are currently limited by manufacturing capabilities and assembly tolerances. Nevertheless, the compact design of brazed aluminum plate–fin exchangers makes it possible even now to furnish more than 35,000 m² of heat-transfer surface in one manifolded assembly. These exchangers are finding application throughout the world in such specific processes as helium liquefaction, helium extraction from natural gas, hydrogen purification and liquefaction,

air separation, and low-temperature hydrocarbon processing. Additional design details for plate–fin exchangers are available in most heat-exchanger texts.

3. Reversing Exchangers

Operation of low-temperature processes on a continuous basis necessitates the removal of all impurities that would solidify on cooling to very low temperatures. This cleanup is necessary because an accumulation of impurities in certain parts of the system can create operational difficulties or constitute potential hazards. Under certain conditions, the necessary purification steps can be carried out with the aid of reversing heat exchangers.

A typical arrangement of a reversing exchanger for an air-separation plant is shown in Fig. 12. Channels A and B constitute the two main reversing streams. Operation of such an exchanger is characterized by the cyclical changeover of one of these streams from one channel to the other. The reversal normally is accomplished by pneumatically operated valves on the warm end and by check valves on the cold end of the exchanger. Feed enters the warm end of the exchanger, and as it is progressively cooled, impurities are deposited on the cold surface of the exchanger. When the flows are reversed, the waste stream reevaporates the deposited impurities and removes them from the system. Pressure differences of the two streams, which in turn affect the saturation concentrations of impurities in those streams, permit impurities to be deposited during the warming period and reevaporated during the cooling period. Temperature differences, particularly at the cold end

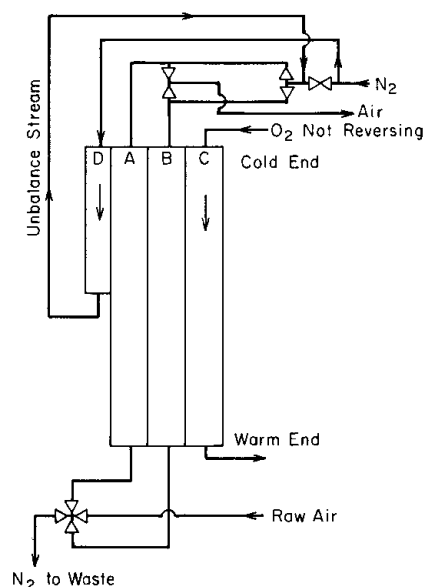


FIGURE 12 Typical flow arrangement for reversing exchanger in air separation plant.

of the reversing heat exchanger, are critical to the proper functioning of these types of exchangers.

4. Regenerators

Another method for the simultaneous cooling and purification of gases in low-temperature processes is based on the use of regenerators, first suggested by Fränkl in the 1920s. Whereas in the reversing exchanger the flows of the two fluids are continuous and countercurrent during any one period, the regenerator operates periodically, storing heat in a high heat-capacity packing in one-half of the cycle and then giving up the stored heat to the fluid in the other half of the cycle.

Such an exchanger normally consists of two identical columns packed with a material of high heat capacity and high heat-transfer area through which the gases that are to be cooled or warmed flow. Such regenerator materials and geometries generally fall into three groups, based on the temperature range over which they are to be used. The first group includes woven screen materials of stainless steel, bronze, or copper used over the temperature range from 30 to 300 K. In the range between 10 and 30 K, lead and antimony spheres are used because their heat capacity is higher than any of the screen materials. However, below 10 K, lead loses 89% of its room-temperature specific heat, and its volumetric heat capacity is less than that of helium at a pressure of 1 MPa. In the late 1980s, a third category of essentially heavy rare-earth intermetallic compounds was developed with the potential for enhancing the heat capacity at temperatures below 10 K. The increase in specific heat of two of these rare-earth compounds is shown in Fig. 13.

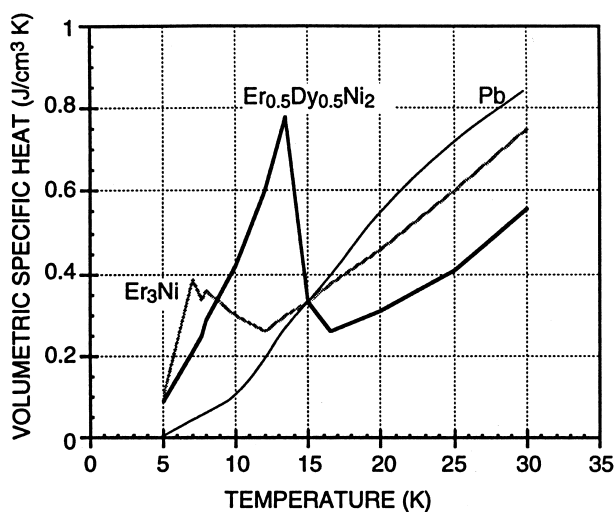


FIGURE 13 Volumetric specific heat of two rare-earth intermetallic compounds and lead.

In the process of cooldown, the warm feed stream deposits impurities on the cold surface of the packing. When the streams are switched, the impurities are reevaporated in the cold stream while simultaneously cooling the packing. Thus, the purifying action of the regenerator is based on the same principles as for the reversing exchanger, and the same limiting critical temperature differences must be observed if complete reevaporation of the impurities is to take place.

Regenerators quite frequently are chosen for applications where the heat-transfer effectiveness, defined as $Q_{\text{actual}}/Q_{\text{ideal}}$, must approach values of 0.98 to 0.99. It is clear that a high regenerator effectiveness requires a high heat capacity per unit volume and a large surface area per unit volume.

The low cost of the heat-transfer surface along with the low pressure drop are the principal advantages of the regenerator. However, the intercontamination of fluid streams by mixing due to periodic flow reversals and the difficulty of regenerator design to handle three or more fluids have restricted its use and favored the adoption of brazed aluminum exchangers.

VI. STORAGE AND TRANSFER SYSTEMS

Once a cryogen has been produced, it must be stored, transferred, or transported to its end use. The effectiveness of the cryogenic storage transfer or transport system depends on how well it reduces the loss of the cryogen due to unavoidable heat leak into the system and how well it maintains the purity of the cryogen. Good design, with a knowledge of the heat-transfer mechanisms and the properties of available insulations, is essential in minimizing the boil-off losses due to heat leak. Proper operating procedures, on the other hand, are necessary if product purity is to be maintained.

A. Insulation Concepts

Since heat leak is a major concern in storage and transfer systems of cryogenic liquids, selecting the proper insulation to use in such systems is vitally important. The normal design strategy is to minimize radiative and convective heat transfer while introducing a minimum of solid conductance media. The choice of insulation, however, is generally governed by an attempt to balance the cost of installed insulation with the savings anticipated by lowered boil-off losses.

The various types of insulation used in the storage and transfer of cryogenic liquids can be divided into five categories: (1) vacuum, (2) multilayer, (3) powder and fibrous, (4) foam, and (5) special. The boundaries between these

TABLE III Characteristics of Selected Insulation

Type of insulation	Apparent thermal conductivity (k_a , J/sec · m · K) (between 77 and 300 K)	Bulk density (kg/m ³)
Pure gas at 0.1 MPa, 180 K		
H ₂	34.07×10^{-2}	0.080
N ₂	5.67×10^{-2}	1.21
Pure vacuum, 0.13 mPa or less	1.70×10^{-2}	Nil
Straight insulation		
Polystyrene foam	8.52×10^{-2}	32–48
Polyurethane foam	10.79×10^{-2}	80–128
Glass foam	11.36×10^{-2}	144
Evacuated powder		
Perlite (133 mPa)	$0.34\text{--}0.68 \times 10^{-2}$	144–64
Silica (133 mPa)	$0.57\text{--}0.68 \times 10^{-2}$	64–96
Combination insulation		
Aluminum foil and fiberglass		
(12–28 layers/cm, 1.33 mPa)	$1.14\text{--}2.27 \times 10^{-4}$	64–112
(30–60 layers/cm, 1.33 mPa)	0.57×10^{-4}	120
Aluminum foil and nylon net		
(32 layers/cm, 1.33 mPa)	5.68×10^{-4}	89

general categories are by no means distinct. However, the classification scheme does offer a framework by which the widely varying types of insulation can be discussed.

Since heat transfer through these insulations can occur by several different mechanisms, the apparent thermal conductivity k_a of an insulation that incorporates all of these heat-transfer possibilities offers the best means of comparing these difference types. Table III provides a listing of some accepted k_a values for popular insulations used in cryogenic storage and transfer systems.

1. Vacuum Insulation

The mechanism of heat transfer prevailing across an evacuated space (0.13 mPa or less) is by radiation and conduction through the residual gas. Radiation is generally the more predominant mechanism and can be approximated by:

$$\frac{Q_r}{A_1} = \sigma(T_2^4 - T_1^4) \left[\frac{1}{\varepsilon_1} + \frac{A_1}{A_2} \left(\frac{1}{\varepsilon_2} - 1 \right) \right]^{-1} \quad (7)$$

where Q_r/A_1 is the radiant heat flux, σ the Stefan-Boltzmann constant, and ε the emissivity of the surface. The subscripts 1 and 2 refer to the cold and warm surfaces, respectively. The bracketed term on the right is generally referred to as the emissivity factor.

When the mean free path of gas molecules becomes large relative to the distance between the walls of the

evacuated space as the pressure is reduced, free molecular conduction is encountered. The gaseous heat conduction under free molecular conditions for most cryogenic applications is given by:

$$\frac{Q_{gc}}{A_1} = \frac{\gamma + 1}{\gamma - 1} \left(\frac{R}{8\pi MT} \right)^{1/2} \alpha p (T_2 - T_1) \quad (8)$$

where α , the overall accommodation coefficient, is defined by:

$$\alpha = \frac{\alpha_1 \alpha_2}{\alpha_2 + \alpha_1(1 - \alpha_2)(A_1/A_2)} \quad (9)$$

and γ is the ratio of the heat capacities, R the molar gas constant, M the molecular weight of the gas, and T the temperature of the gas at the point where the pressure p is measured. The subscripted A_1 and A_2 , T_1 and T_2 , and α_1 and α_2 are the areas, temperatures, and accommodation coefficients of the cold and warm surfaces, respectively. The accommodation coefficient depends on the specific gas-surface combination and the surface temperature.

Heat transport across an evacuated space by radiation can be reduced significantly by inserting one or more low-emissivity floating shields within the evacuated space. Such shields provide a reduction in the emissivity factor. The only limitation on the number of floating shields used is one of complexity and cost.

2. Multilayer Insulation

Multilayer insulation provides the most effective thermal protection available for cryogenic storage and transfer systems. It consists of alternating layers of highly reflecting material, such as aluminum foil or aluminized Mylar, and a low-conductivity spacer material or insulator, such as fiberglass mat or paper, glass fabric, or nylon net, all under high vacuum. When properly applied at the optimum density, this type of insulation can have an apparent thermal conductivity as low as 10 to 50 $\mu\text{W/m} \cdot \text{K}$ between 20 and 300 K. The very low thermal conductivity of multilayer insulations can be attributed to the fact that all modes of heat transfer are reduced to a bare minimum.

The apparent thermal conductivity of a highly evacuated (pressures on the order of 0.13 mPa or less) multilayer insulation can be determined from:

$$k_a = \frac{1}{N/\Delta x} \left[h_s + \frac{\sigma e T_2^3}{2 - e} \left(1 + \frac{T_1}{T_2} \right)^2 \left(1 + \frac{T_1}{T_2} \right) \right] \quad (10)$$

where $N/\Delta x$ is the number of complete layers (reflecting shield plus spacer) of insulation per unit thickness, h_s the solid conductance for the spacer material, σ the Stefan-Boltzmann constant, e the effective emissivity of the reflecting shield, and T_2 and T_1 the temperatures of

the warm and cold sides of the insulation, respectively. It is evident that the apparent thermal conductivity can be reduced by increasing the layer density up to a certain point.

Unfortunately, the effective thermal conductivity values generally obtained with actual cryogenic storage and transfer systems are at least a factor of 2 greater than the thermal conductivity values measured in the laboratory with carefully controlled techniques. This degradation in insulation thermal performance is caused by the combined presence of edge exposure to isothermal boundaries, gaps, joints, or penetrations in the insulation blanket required for structural supports, fill and vent lines, and the high lateral thermal conductivity of these insulation systems.

3. Powder Insulation

The difficulties encountered with the use of multilayer insulation for complex structural storage and transfer systems can be minimized by the use of evacuated powder insulation. This substitution in insulation materials, however, incurs a 10-fold decrease in overall thermal effectiveness of the insulation system. Nevertheless, in applications where this is not a serious factor and investment cost is a major factor, even unevacuated powder insulation with still a lower thermal effectiveness may be the proper choice of insulating material. Such is the case for large LNG storage facilities.

A powder insulation system consists of a finely divided particulate material such as perlite, expanded SiO_2 , calcium silicate, diatomaceous earth, or carbon black packed between the surfaces to be insulated. When used at 0.1 MPa gas pressure (generally with an inert substance), the powder reduces both convection and radiation and, if the particle size is sufficiently small, can also reduce the mean free path of the gas molecules.

The radiation contribution for highly evacuated powders near room temperature is larger than the solid-conduction contribution to the total heat transfer rate. On the other hand, the radiant contribution is smaller than the solid-conduction contribution for temperatures between 77 and 20 or 4 K. Thus, evacuated powders can be superior to vacuum alone (for insulation thicknesses greater than ~ 0.1 m) for heat transfer between ambient and liquid nitrogen temperatures. Conversely, since solid conduction becomes predominant at lower temperatures, it is usually more advantageous to use vacuum alone for reducing heat transfer between two cryogenic temperatures.

4. Foam Insulation

The apparent thermal conductivity of foams is dependent on the bulk density of the foamed material, the gas used

as the foaming agent, and the temperature levels to which the insulation is exposed. Heat transport across a foam is determined by convection and radiation within the cells of the foam and by conduction in the solid structure. Evacuation of a foam is effective in reducing its thermal conductivity, indicating a partially open cellular structure, but the resulting values are still considerably higher than either multilayer or evacuated powder insulations. The opposite effect, diffusion of atmospheric gases into the cells, can cause an increase in the apparent thermal conductivity. This is particularly true with the diffusion of hydrogen and helium into the cells. Of all the foams, polyurethane and polystyrene have received the widest use at low temperatures.

The major disadvantage of foams is not their relatively high thermal conductivity compared with that of other insulations, but rather their poor thermal behavior. When applied to cryogenic systems, they tend to crack on repeated cycling and lose their insulation value.

5. Special Insulations

No single insulation has all the desirable thermal and strength characteristics required in many cryogenic applications. Consequently, numerous composite insulations have been developed. One such insulation consists of a polyurethane foam, reinforcement of the foam to provide adequate compressive strength, adhesives for sealing and securing the foam to the container, enclosures to prevent damage to the foam from external sources, and vapor barriers to maintain a separation between the foam and atmospheric gases. Another external insulation system for space applications uses honeycomb structures. Phenolic resin-reinforced fiberglass-cloth honeycomb is most commonly used. Filling the cells with a low-density polyurethane foam further improves the thermal effectiveness of the insulation.

B. Storage Systems

Storage vessels range from low-performance containers where the liquid in the container boils away in a few hours to high-performance containers and dewars where less than 0.1% of the fluid contents is evaporated per day. Since storage and transfer systems are important components of any cryogenic support facility, many examples of storage vessel design have appeared in the literature. The essential elements of a storage vessel consist of an inner vessel, which encloses the cryogenic fluid to be stored, and an outer vessel, which contains the appropriate insulation and serves as a vapor barrier to prevent water and other condensables from reaching the cold inner vessel. The value of the cryogenic liquid stored will dictate whether or not

the insulation space is evacuated. In small laboratory dewars, the insulation is obtained by coating the two surfaces facing the insulation space with low-emissivity materials and then evacuating the space to a pressure of 0.13 mPa or lower. In larger vessels, insulations such as powders, fibrous materials, or multilayer insulations are used.

Several requirements must be met in the design of the inner vessel. The material of construction must be compatible with the stored cryogen. Nine percent nickel steels are acceptable for high-boiling cryogenics ($T > 75$ K), while many aluminum alloys and austenitic steels are usually structurally acceptable throughout the entire temperature range. Economic and cooldown considerations dictate that the inner shell be as thin as possible. Accordingly, the inner container is designed to withstand only the internal pressure and bending forces, while stiffening rings are used to support the weight of the fluid. The minimum thickness of the inner shell for a cylindrical vessel under such a design arrangement is given in Section VIII of the American Society of Mechanical Engineers' (ASME) *Boiler and Pressure Vessel Code*.

The outer shell of the storage vessel, on the other hand, is subjected to atmospheric pressure on the outside and evacuated conditions on the inside. Such a pressure difference requires an outer shell of sufficient material thickness with appropriately placed stiffening rings to withstand collapsing or buckling. Here again, specific design charts addressing this situation can be found in the ASME code.

Heat leak into a storage system for cryogenics generally occurs by radiation and conduction through the insulation and conduction through the inner shell supports, piping, instrumentation leads, and access ports. Conduction losses are reduced by introducing long heat-leak paths, by making the cross-sections for heat flow small, and by using materials with low thermal conductivity. Radiation losses, a major factor in the heat leak through insulations, are reduced with the use of radiation shields, such as multilayer insulation, boil-off vapor-cooled shields, and opacifiers in powder insulation.

Most storage vessels for cryogenics are designed for a 90% liquid volume and a 10% vapor or ullage volume. The latter permits reasonable vaporization of the liquid contents due to heat leak without incurring too rapid a buildup of pressure in the vessel. This, in turn, permits closure of the container for short periods either to avoid partial loss of the contents or to permit the safe transport of flammable or hazardous cryogenics.

C. Transfer Systems

Three methods are commonly used to transfer a cryogen from the storage vessel. These are self-pressurization of the container, external gas pressurization, and mechanical

pumping. Self-pressurization involves removing some of the fluid from the container, vaporizing the extracted fluid, and then reintroducing the vapor into the ullage space, thereby displacing the contents of the container. The external gas pressurization method utilizes an external gas to accomplish the desired displacement of the container contents. In the mechanical pumping method, the contents of the storage vessel are removed by a cryogenic pump located in the liquid drain line.

Several different types of pumps have been used with cryogenic fluids. In general, the region of low flow rates at high pressures is best suited to positive displacement pumps, while the high-flow applications are generally best served by the use of centrifugal or axial flow pumps. The latter have been built and used for liquid hydrogen with flow rates of up to 3.8 m³/sec and pressures of more than 6.9 MPa. For successful operation, cryogen subcooling, thermal contraction, lubrication, and compatibility of materials must be carefully considered.

Cryogenic fluid transfer lines are generally classified as one of three types: uninsulated, foam-insulated lines, and vacuum-insulated lines. The latter may entail vacuum insulation alone, evacuated powder insulation, or multilayer insulation. A vapor barrier must be applied to the outer surface of foam-insulated transfer lines to minimize the degradation of the insulation that occurs when water vapor and other condensables are permitted to diffuse through the insulation to the cold surface of the lines.

Two-phase flow is always involved in the cooldown of a transfer line. Since this process is a transient one, several different types of two-phase flow will exist simultaneously along the inlet of the transfer line. Severe pressure and flow oscillations occur as the cold liquid comes in contact with successive warm sections of the line. Such instability continues until the entire transfer line is cooled down and filled with liquid cryogen.

The transport of cryogenics for more than a few hundred meters generally requires specially built transport systems for truck, railroad, or airline delivery. Volumes from 0.02 to more than 100 m³ have been transported successfully by these carriers. The use of large barges and ships built specifically for cryogen shipment has increased the volume transported manyfold. This has been particularly true for the worldwide transport of LNG.

VII. INSTRUMENTATION

Once low temperatures have been attained and cryogenics have been produced, property measurements must often be made at these temperatures. Such measurements as temperature and pressure are typically required for process optimization and control. In addition, as cryogenic fluids

have acquired greater commercial importance, questions have arisen relative to the quantities of these fluids transferred or delivered. Accordingly, the instrumentation used must be able to indicate liquid level, density, and flow rate accurately.

A. Thermometry

Most low-temperature engineering temperature measurements are made with metallic resistance thermometers, nonmetallic resistance thermometers, or thermocouples. In the selection of a thermometer for a specific application one must consider such factors as absolute accuracy, reproducibility, sensitivity, heat capacity, self-heating, heat conduction, stability, simplicity and convenience of operation, ruggedness, and cost. Other characteristics may be of importance in certain applications.

B. Fluid Measurements

Liquid level is one of several measurements needed to establish the contents of a cryogenic container. Other measurements may include volume as a function of depth, density as a function of physical storage conditions, and sometimes discerning useful contents from total contents. Of these measurements, the liquid-level determination is presently the most advanced and can be made with an accuracy and precision comparable to that of thermometry and often with greater simplicity.

There are as many ways of classifying liquid-level sensors as there are developers who have described them. A convenient way to classify such devices is according to whether the output is discrete (point sensors) or continuous.

C. Density Measurements

Measurements of liquid density are closely related to quantity and liquid-level measurements since both are often required simultaneously to establish the mass contents of a tank, and the same physical principle may often be used for either measurement, since liquid-level detectors sense the steep density gradient at the liquid-vapor interface. Thus, the methods of density determination include the following techniques: direct weighing, differential pressure, capacitance, optical, acoustic, and nuclear radiation attenuation. In general, the various liquid level principles apply to density measurement techniques as well.

Two exceptions are noteworthy. In the case of homogeneous pure fluids, density can usually be determined more accurately by an indirect measurement, namely, the measurement of pressure and temperature which is then coupled with the analytical relationship between these in-

tensive properties and density through accurate thermophysical properties data.

The case of nonhomogeneous fluids is quite different. LNG is often a mixture of five or more components whose composition and, hence, density vary with time and place. Accordingly, temperature and pressure measurements alone will not suffice. A dynamic, direct measurement is required, embodying one or more of the liquid-level principles used in liquid-level measurements.

D. Flow Measurements

Three basic types of flow meters are useful for liquid cryogenics. These are the pressure drop or "head" type, the turbine type, and the momentum type.

VIII. SAFETY

No discussion of cryogenic systems would be complete without a review of some of the safety aspects associated with either laboratory or industrial use of cryogenic fluids. Earlier discussion of the properties of cryogenic fluids and the behavior of materials at low temperatures revealed that there are a number of unique hazards associated with cryogenic fluids. These hazards can best be classified as those associated with the response of the human body and the surroundings to cryogenic fluids and their vapors, and those associated with reactions between certain of the cryogenic fluids and their surroundings.

A. Human Hazards

It is well known that exposure of the human body to cryogenic fluids or to surfaces cooled by cryogenic fluids can result in severe "cold burns" since damage to the skin or tissue is similar to that caused by an ordinary burn. The severity of the burn depends on the contact area and the contact time; prolonged contact results in deeper burns. Severe burns are seldom sustained if rapid withdrawal is possible.

Protective clothing is mandatory to insulate the body from these low temperatures and prevent "frostbite." Safety goggles, gloves, and boots are imperative for personnel involved in the transfer of liquid cryogenics. Such transfers, in the interest of good safety practices, should be attempted only when sufficient personnel are available to monitor the activity. Since nitrogen is a colorless, odorless, inert gas, personnel must be aware of the associated respiratory and asphyxiation hazards. Whenever the oxygen content of the atmosphere is diluted due to spillage or leakage of nitrogen, there is danger of nitrogen asphyxiation. In general, the oxygen content of air for breathing

purposes should never be below 16%. Whenever proper air ventilation cannot be ensured, air-line respirators or a self-contained breathing apparatus should be used.

An oxygen-enriched atmosphere, on the other hand, produces exhilarating effects when breathed. However, lung damage can occur if the oxygen concentration in the air exceeds 60%, and prolonged exposure to an atmosphere of pure oxygen may initiate bronchitis, pneumonia, or lung collapse. An additional threat of oxygen-enriched air can come from the increased flammability and explosion hazards.

B. Materials Compatibility

Most failures of cryogenic systems can generally be traced to an improper selection of construction materials or a disregard for the change of some material property from ambient to low temperatures. For example, the ductility property of a material requires careful consideration since low temperatures have the effect of making some construction materials brittle or less ductile. This behavior is further complicated because some materials become brittle at low temperatures but still can absorb considerable impact, while others become brittle and lose their impact strength. Brittle fracture can occur very rapidly, resulting in almost instantaneous failure. Such failure can cause shrapnel damage if the system is under pressure, while release of a fluid such as oxygen can result in fire or explosions.

Low-temperature equipment can also fail because of thermal stresses caused by thermal contraction of the materials used. In solder joints, the solder must be able to withstand stresses caused by differential contraction where two dissimilar metals are joined. Contraction in long pipes is also a serious problem; a stainless-steel pipeline 30 m long will contract ~ 0.085 m when filled with liquid oxygen or nitrogen. Provisions must be made for this change in length during both cooling and warming of the pipeline by using bellows, expansion joints, or flexible hose. Pipe anchors, supports, and so on likewise must be carefully designed to permit contraction and expansion to take place. The primary hazard of failure due to thermal contraction is spillage of the cryogen and the possibility of fire or explosion.

All cryogenic systems should be protected against overpressure due to phase change from liquid to gas. Systems containing liquid cryogens can reach bursting pressures, if not relieved, simply by trapping the liquid in an enclosure. The rate of pressure rise depends on the rate of heat transfer into the liquid. In uninsulated systems, the liquid is vaporized rapidly and pressure in the closed system can rise very rapidly. The more liquid there is originally in the tank before it is sealed off, the greater will be the

resulting final pressure. Relief valves and burst disks are normally used to relieve piping systems at a pressure near the design pressure of the equipment. Such relief should be provided between valves, on tanks, and at all points of possible (though perhaps unintentional) pressure rise in a piping system.

Overpressure in cryogenic systems can also occur in a more subtle way. Vent lines without appropriate rain traps can collect rainwater. Which when frozen can block the line. Exhaust tubes on relief valves and burst disks likewise can become inoperable. Small-necked, open-mouth dewars can collect moisture from the air and freeze closed. Entrapment of cold liquids or gases can occur by freezing water or other condensables in some portion of the cold system. If this occurs in an unanticipated location, the relief valve or burst disk may be isolated and afford no protection. Such a situation usually arises from improper operating procedures and emphasizes the importance of good operating practices.

Another source of system overpressure that is frequently overlooked results from cooldown surges. If a liquid cryogen is admitted to a warm line for the purpose of transfer of the liquid from one point to another, severe pressure surges will occur. These pressure surges can be up to 10 times the operating or transfer pressure and can even cause backflow into the storage container. Protection against such overpressure must be included in the overall design and operating procedures for the transfer system.

In making an accident or safety analysis, it is always wise to consider the possibility of encountering even more serious secondary effects from any cryogenic accident. For example, any one of the failures discussed previously (brittle fracture, contraction, overpressure, etc.) may release sizable quantities of cryogenic liquids, causing a severe fire or explosion hazard, asphyxiation possibilities, further brittle fracture problems, or shrapnel damage to other flammable or explosive materials. In this way the situation can rapidly and progressively become much more serious.

C. Flammability and Detonability

Almost any flammable mixture will, under favorable conditions of confinement, support an explosive flame propagation or even a detonation. When a fuel-oxidant mixture of a composition favorable for high-speed combustion is weakened by dilution with an oxidant, fuel, or an inert substance, it will first lose its capacity to detonate. Further dilution will then cause it to lose its capacity to burn explosively. Eventually, the lower or upper flammability limits will be reached and the mixture will not maintain its combustion temperature and will automatically extinguish itself. These principles apply to the combustible cryogens hydrogen and methane. The flammability and detonability

TABLE IV Flammability and Detonability Limits of Hydrogen and Methane Gas

Mixture	Flammability limits (mol%)	Detonability limits (mol%)
H ₂ -air	4-75	20-65
H ₂ -O ₂	4-95	15-90
CH ₄ -air	5-15	6-14
CH ₄ -O ₂	5-61	10-50

limits for these two cryogens with either air or oxygen are presented in Table IV. Since the flammability limits are rather broad, great care must be exercised to exclude oxygen from these cryogens. This is particularly true with hydrogen since even trace amounts of oxygen will condense, solidify, and build up with time in the bottom of the liquid hydrogen storage container and eventually attain the upper flammability limits. Then it is just a matter of time until some ignition source, such as a mechanical or electrostatic spark, accidentally initiates a fire or possibly an explosion.

Because of its chemical activity, oxygen also presents a safety problem in its use. Liquid oxygen is chemically reactive with hydrocarbon materials. Ordinary hydrocarbon lubricants are even dangerous to use in oxygen compressors and vacuum pumps exhausting gaseous oxygen. In fact, valves, fittings, and lines used with oil-pumped gases should never be used with oxygen. Serious explosions have resulted from the combination of oxygen and hydrocarbon lubricants.

To ensure against such unwanted chemical reactions, systems using liquid oxygen must be kept scrupulously clean of any foreign matter. The phrase "LOX clean" in the space industry has come to be associated with a set of elaborate cleaning and inspection specifications nearly representing the ultimate in large-scale equipment cleanliness.

Liquid oxygen equipment must also be constructed of materials incapable of initiating or sustaining a reaction. Only a few polymeric materials can be used in the design of such equipment since most will react violently with oxygen under mechanical impact. Also, reactive metals such as titanium and aluminum should be used cautiously, since they are potentially hazardous. Once the reaction is started, an aluminum pipe containing oxygen burns rapidly and intensely. With proper design and care, however, liquid oxygen systems can be operated safely.

Even though nitrogen is an inert gas and will not support combustion, there are some subtle means whereby a flammable or explosive hazard may develop. Cold traps or open-mouth dewars containing liquid nitrogen can con-

dense air and cause oxygen enrichment of the liquid nitrogen. The composition of air as it condenses into the liquid nitrogen container is about 50% oxygen and 50% nitrogen. As the liquid nitrogen evaporates, the liquid oxygen content steadily increases so that the last portion of liquid to evaporate will have a relatively high oxygen concentration. The nitrogen container must then be handled as if it contained liquid oxygen. Explosive hazards all apply to this oxygen-enriched liquid nitrogen.

Since air condenses at temperatures below ~82 K, uninsulated pipelines transferring liquid nitrogen will condense air. This oxygen-enriched condensate can drip on combustible materials, causing an extreme fire hazard or explosive situation. The oxygen-rich air condensate can saturate clothing, rags, wood, asphalt pavement, and so on and cause the same problems associated with the handling and spillage of liquid oxygen.

D. Summary

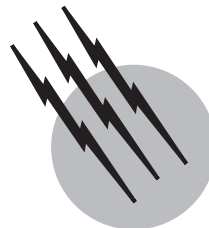
It is obvious that the best designed cryogenic facility is no better than the attention paid to every potential hazard. Unfortunately, the existence of such potential hazards cannot be considered once and then forgotten. Instead, there must be an ongoing safety awareness that focuses on every conceivable hazard that might be encountered. Assistance with identifying these safety hazards is adequately covered by Edeskuty and Stewart (1996).

SEE ALSO THE FOLLOWING ARTICLES

CHEMICAL ENGINEERING THERMODYNAMICS • CRYOGENICS • HEAT EXCHANGERS • METALLURGY, MECHANICAL • SUPERCONDUCTIVITY MECHANISMS • VACUUM TECHNOLOGY

BIBLIOGRAPHY

- Barron R. F. (1986). "Cryogenic Systems," Oxford Univ. Press, London.
- Edeskuty, F. J., and Stewart, W. F. (1996). "Safety in the Handling of Cryogenic Fluids," Plenum Press, New York.
- Flynn, T. M. (1996). "Cryogenic Engineering," Dekker, New York.
- Jacobsen, R. T., Penoncello, S. G., and Lemmon, E. W. (1997). "Thermodynamic Properties of Cryogenic Fluids," Plenum Press, New York.
- Ross, R. G., Jr. (1999). "Cryocoolers 10," Kluwer Academic/Plenum Publishers, New York.
- Timmerhaus, K. D., and Flynn, T. M. (1989). "Cryogenic Process Engineering," Plenum Press, New York.
- Van Sciver, S. W. (1986). "Helium Cryogenics," Plenum Press, New York.
- Weisend, J. G., II (1998). "Handbook of Cryogenic Engineering," Taylor & Francis, London.



Crystallization Processes

Ronald W. Rousseau

Georgia Institute of Technology

- I. Objectives of Crystallization Processes
- II. Equilibrium and Mass and Energy Balances
- III. Nucleation and Growth Kinetics
- IV. Purity, Morphology, and Size Distributions
- V. Crystallizer Configuration and Operation
- VI. Population Balances and Crystal Size Distributions

GLOSSARY

Crystallizer The vessel or process unit in which crystallization occurs.

Growth The increase in crystal size due to deposition of solute on crystal surfaces.

Magma The mixture of crystals and mother liquor in the crystallizer.

Mode of crystallization The means by which a thermodynamic driving force for crystallization is created.

Mother liquor The liquid solution from which crystals are formed.

MSMPR crystallizer A vessel operating in a continuous manner in which crystallization occurs and whose contents are perfectly mixed. As a result of perfect mixing, all variables descriptive of the mother liquor and crystals are constant throughout the vessel and are identical to corresponding variables in the product stream leaving the vessel.

Nucleation The formation of new crystals.

Primary nucleation The formation of crystals by mechanisms that do not involve existing crystals of the crys-

tallizing species; includes both homogeneous and heterogeneous nucleation mechanisms.

Secondary nucleation The formation of crystals through mechanisms involving existing crystals of the crystallizing species.

Solubility The equilibrium solute concentration. The dimensions in which solubility is expressed include, but are not limited to, mass or mole fraction, mass or mole ratio of solute to solvent, and mass or moles of solute per unit volume of solvent or solution.

Supersaturation The difference between existing and equilibrium conditions; the quantity represents the driving force for crystal nucleation and growth.

CRYSTALLIZATION PROCESSES addressed in this discussion are used in the chemical, petrochemical, pharmaceutical, food, metals, agricultural, electronics, and other industries. Moreover, the principles of crystallization are important in all circumstances in which a solid crystalline phase is produced from a fluid, even when the solid is not a product of the process. Much has been done

in recent years to improve the understanding of crystallization, and a large portion of the research on the topic has dealt with mechanisms of nucleation and growth. Especially important has been elucidation of the effects of process variables on the rates at which these phenomena occur. Additionally, extensive progress has been achieved in modeling both steady-state and dynamic behavior of crystallization systems of industrial importance. The primary elements of the discussion that follows are the principles that influence yield, morphology, and size distribution of crystalline products.

I. OBJECTIVES OF CRYSTALLIZATION PROCESSES

Several examples of objectives that may be satisfied in crystallization processes are given in the following discussion. Soda ash (sodium carbonate) is recovered from brine by contacting the brine with carbon dioxide that reacts with sodium carbonate to form sodium bicarbonate. Sodium bicarbonate, which has a lower solubility than sodium carbonate, crystallizes as it is formed. The primary objective of the crystallizers used in this process is separation of a high percentage of sodium bicarbonate from the brine in a form that facilitates segregation of the crystals from the mother liquor. The economics of crystal separation from the mother liquor are affected primarily by the variables that control the flow of liquid through the cake of crystals formed on a filter or in a centrifuge. For example, the flow rate of liquid through a filter cake depends on the available pressure drop across a filter, liquid viscosity, and the size distribution of crystals collected on the filter. With a fixed available pressure drop and defined liquid properties, the crystal size distribution controls filter throughput and, concomitantly, the production rate from the process.

Crystallization can be used to remove solvent from a liquid solution. For example, concentration of fruit juice requires the separation of solvent (water) from the natural juice. The common procedure is evaporation, but the derived juices may lose flavor components or undergo thermal degradation during the evaporative process. In freeze concentration, the solvent is crystallized (frozen) in relatively pure form to leave behind a solution with a higher solute concentration than the original mixture. Significant advantages in product taste have been observed in the application of this process to concentrations of various types of fruit juice.

The elimination of small amounts of an impurity from a product species may be an objective of crystallization. In such instances, a multistep crystallization–redissolution–recrystallization process may be required to produce a

product that meets purity specifications. For example, in the manufacture of the amino acid L-isoleucine, the product is first recovered in acid form, redissolved, neutralized, and then recrystallized in order to exclude the impurity L-leucine and other amino acids from the product.

A simple change in physical properties also can be achieved by crystallization. In the process of making soda ash, referred to earlier, the sodium bicarbonate crystals are subjected to heat that causes the release of carbon dioxide and produces low-density sodium carbonate crystals. The density of these crystals is incompatible with their use in glass manufacture, but a more acceptable crystal can be obtained by contacting the sodium carbonate crystals with water to form crystalline sodium carbonate monohydrate. Drying the resulting crystals removes the water of hydration and produces a dense product that is acceptable for glass manufacture.

Separation of a chemical species from a mixture of similar compounds can be achieved by crystallization. The mode of crystallization may fall in the realm of what is known as melt crystallization. In such processes, the mother liquor largely is comprised of the melt of the crystallizing species, and, subsequent to its crystallization, crystals formed from the mother liquor are remelted to produce the product from the crystallizer. *Para(p)*-xylene can be crystallized from a mixture that includes *ortho* and *meta* isomers in a vertical column that causes crystals and mother liquor to move countercurrently. Heat is added at the bottom of the column to melt the *p*-xylene crystals; a portion of the melt is removed from the crystallizer as product and the remainder flows up the column to contact the downward-flowing crystals. Effluent mother liquor, consisting almost entirely of the *ortho* and *meta* isomers of xylene, is removed from the top of the column.

Production of a consumer product in a form suitable for use and acceptable to the consumer also may be an objective of a crystallization process. For example, sucrose (sugar) can be crystallized in various forms. However, different cultures are accustomed to using sugar that has a particular appearance and, unless the commodity has that appearance, the consumer may consider the sugar to be unacceptable.

In all these processes, there are common needs: to form crystals, to cause them to grow, and to separate the crystals from the residual liquid. While conceptually simple, the operation of a process that utilizes crystallization can be very complex. The reasons for such complexity involve the interaction of the common needs and process requirements on product yield, purity, and, uniquely, crystal morphology and size distribution. In the following discussion, the interactions will be discussed and general principles affecting crystallizer operation will be outlined. More

extensive discussion of the subject matter can be found in the bibliography at the end of the chapter.

II. EQUILIBRIUM AND MASS AND ENERGY BALANCES

A. Solid–Liquid Equilibrium

The solubility of a chemical species in a solvent refers to the amount of solute that can be dissolved at constant temperature, pressure, and solvent composition (including the presence of other solutes). In other words, it is the concentration of the solute in the solvent at equilibrium.

As with all multiphase systems, the Gibbs phase rule provides a useful tool for determining the number of intensive variables (ones that do not depend on system mass) that can be fixed independently:

$$N_{DF} = N_c - N_p + 2 \quad (1)$$

N_{DF} is the number of degrees of freedom, N_c is the number of components, and N_p is the number of phases in the system. The number of degrees of freedom represents the number of independent variables that must be specified in order to fix the condition of the system. For example, the Gibbs phase rule specifies that a two-component, two-phase system has two degrees of freedom. If temperature and pressure are selected as the specified variables, then all other intensive variables—in particular, the composition of each of the two phases—are fixed, and solubility diagrams of the type shown for a hypothetical mixture of R and S in Fig. 1 can be constructed.

Several features of the hypothetical system described in Fig. 1 illustrate the selection of crystallizer operating

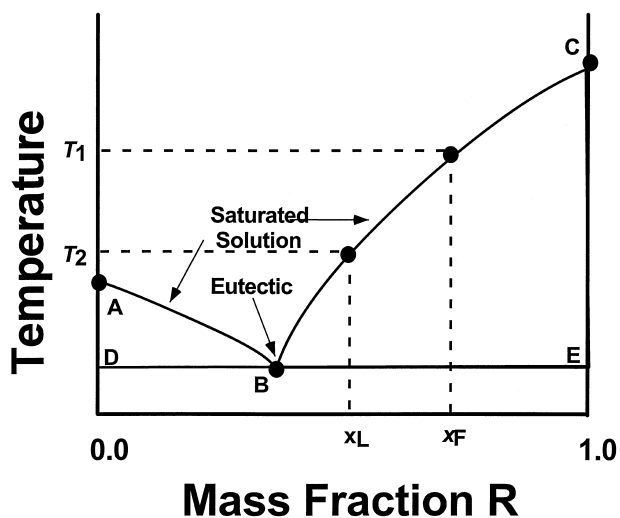


FIGURE 1 Hypothetical solubility diagram of eutectic-forming system.

conditions and the limitations placed on the operation by the system properties. The curves AB and BC represent solution compositions that are in equilibrium with solids whose compositions are given by the lines AD and CE, respectively. If AD and CE are vertical and coincident with the left and right extremes, the crystals are pure S and R, respectively. Crystallization from any solution whose equilibrium composition is to the left of a vertical line through point B will produce crystals of pure S, while solutions with an equilibrium composition to the right of the line will produce crystals of pure R. A solution whose composition falls on the line through B will produce a mixture of crystals of R and S.

Now suppose a saturated solution at temperature T_1 is fed to a crystallizer operating at temperature T_2 . Since it is saturated, the feed has a mole fraction of R equal to x_F . The maximum production rate of crystals occurs when the solution leaving the crystallizer is saturated, meaning that the crystal production rate, m_{prod} , depends on the value of T_2 :

$$m_{\text{prod}} = m_F x_F - m_L x_L \quad (2)$$

where m_F is the feed rate to the crystallizer and m_L is the solution flow rate leaving the crystallizer. Note that the lower limit on T_2 is given by the eutectic point, and that attempts to operate the crystallizer at a temperature other than the eutectic value will result in a mixture of crystals of R and S.

When certain solutes crystallize from aqueous solutions, the crystals are hydrated salts, which means that the crystals contain water and solute in a specific stoichiometric ratio. The water in such instances is referred to as *water of hydration*, and the number of water molecules associated with each solute molecule may vary with the crystallization temperature.

Potassium sulfate provides an example of such behavior. When it crystallizes from an aqueous solution above 40°C, the crystals are anhydrous K_2SO_4 , while below 40°C each molecule of K_2SO_4 that crystallizes has 10 molecules of water associated with it. The hydrated salt, $\text{K}_2\text{SO}_4 \cdot 10\text{H}_2\text{O}(\text{s})$, is called potassium sulfate decahydrate. Another solute that forms hydrated salts is magnesium sulfate, which can incorporate differing amounts of water depending upon the temperature at which crystallization occurs (see Table I).

The solubility diagrams of several species are shown in Fig. 2, and these illustrate the importance of solubility behavior in the selection of the mode of crystallization. For example, consider the differences between potassium nitrate and sodium chloride: The solubility of potassium nitrate is strongly influenced by the system temperature, whereas the opposite is true for sodium chloride. As a consequence, (1) a high yield of potassium nitrate crystals can be obtained by cooling a saturated feed solution,

TABLE I Water of Hydration for MgSO_4

Form	Name	wt% MgSO_4	Conditions
MgSO_4	Anhydrous magnesium sulfate	0.0	$>100^\circ\text{C}$
$\text{MgSO}_4 \cdot \text{H}_2\text{O}$	Magnesium sulfate monohydrate	87.0	67 to 100°C
$\text{MgSO}_4 \cdot 6\text{H}_2\text{O}$	Magnesium sulfate hexahydrate	52.7	48 to 67°C
$\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$	Magnesium sulfate heptahydrate	48.8	2 to 48°C
$\text{MgSO}_4 \cdot 12\text{H}_2\text{O}$	Magnesium sulfate dodecahydrate	35.8	-4 to 2°C

but (2) cooling a saturated sodium chloride solution accomplishes little crystallization, and vaporization of water is required to increase the yield.

The effect of water of hydration on solubility can be seen in Fig. 2. Note, for example, that sodium sulfate has two forms in the temperature range of the solubility diagram: sodium sulfate decahydrate ($\text{Na}_2\text{SO}_4 \cdot 10\text{H}_2\text{O}$), which is known as Glauber's salt, and anhydrous sodium sulfate. Since a transition from Glauber's salt to anhydrous sodium sulfate occurs at approximately 34°C , crystals recovered from a crystallizer operating above about 34°C will be anhydrous, but those from a crystallizer operating below this temperature will contain 10 waters of hydration. Also observe the effect of water of hydration on solubility characteristics; clearly, cooling crystallization could be used to recover significant yields of Glauber's salt but evaporative crystallization would be required to obtain high yields of the anhydrous salt.

Mixtures of multiple solutes in a single solvent are encountered in a number of processes—for example, in the recovery of various chemicals from ores or brines. Express-

sion of the complex solubility behavior in such systems by graphical means usually is limited to systems of two solutes. The interaction of added solutes on solubility is illustrated by the plot of equilibrium behavior for potassium nitrate–sodium nitrate–water in Fig. 3. As before, the curves in the diagram trace solution compositions that are in equilibrium with solid solutes. Points A, D, G, and J are based on the solubilities of pure potassium nitrate, while points C, F, I, and L are based on solubilities of pure sodium nitrate. Curves AB, DE, GH, and JK represent compositions of solutions in equilibrium with solid potassium nitrate at 30, 50, 70, and 100°C , respectively. Curves BC, EF, HI, and KL represent compositions of solutions in equilibrium with solid sodium nitrate. Should the solution condition, including temperature, correspond to points B, E, H, K or any condition on the curve connecting these points, crystals of both solutes would be formed by cooling.

A second type of solubility behavior is exhibited by mixtures that form solid solutions. Consider, for example, a hypothetical system containing R and S whose

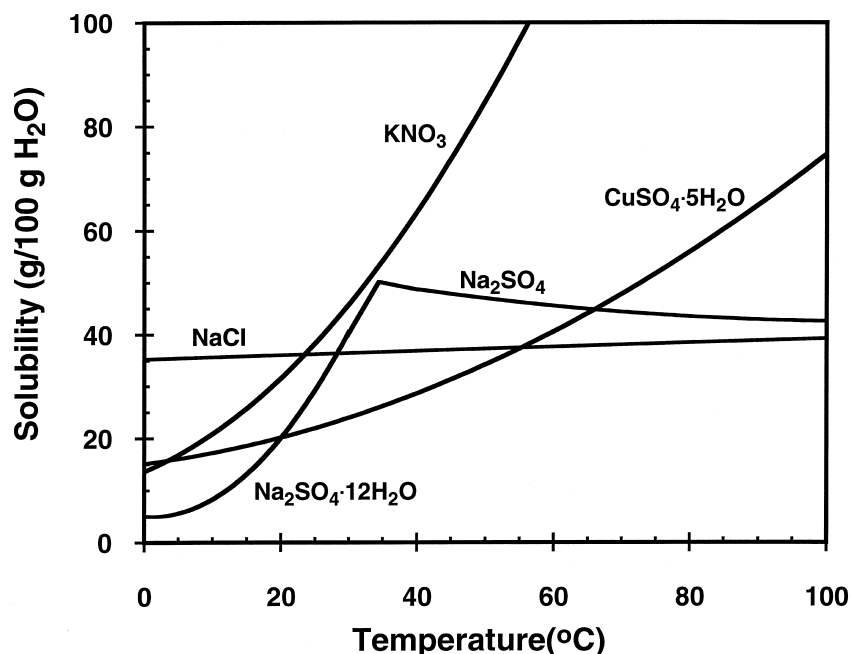


FIGURE 2 Solubility diagram for several common substances.

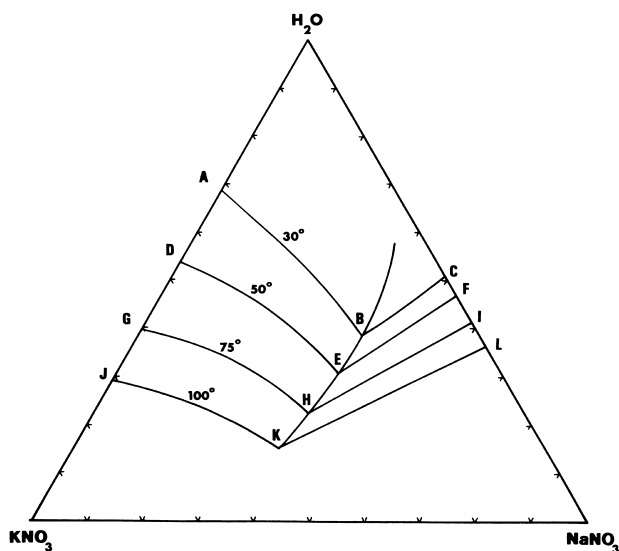


FIGURE 3 Solubility diagram of KNO_3 and NaNO_3 mixtures in water.

equilibrium behavior is described in Fig. 4. The phase envelope is drawn based on the compositions of coexisting liquid and solid phases at equilibrium. The pure component R has a melting point at pressure P equal to T_2 while the melting point of pure S is T_1 . The system behavior can best be described by the following example: Consider a mixture of R and S at temperature T_A and having a mass fraction of R equal to z_M . From the phase diagram, the mixture is a liquid. As the liquid is cooled, a solid phase forms when the temperature reaches T_B and the system is allowed to come to equilibrium; the solid-phase composition corresponds to a mass fraction of R equal to x_B . On cooling the liquid further, the ratio of solid to liquid in-

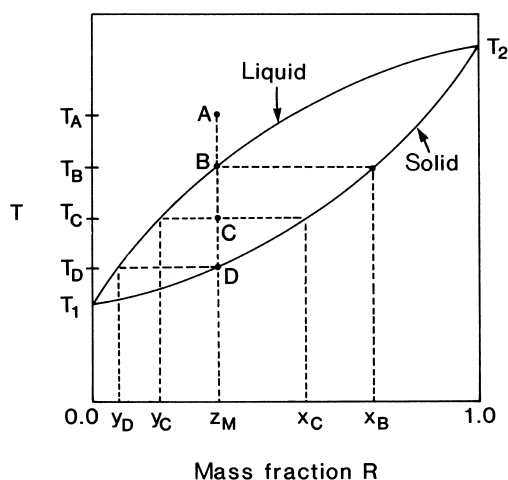


FIGURE 4 Hypothetical solubility diagram of mixture without a eutectic at constant pressure: x , solid; y , liquid; z , combined.

creases and at T_C the mass fraction of R in the liquid is y_C and in the solid it is x_C . At T_D the liquid phase disappears, leaving a solid with a mass fraction of R equal to z_M .

Systems that exhibit behavior of the type illustrated in Fig. 4 cannot be purified in a single crystallization stage. They represent situations in which multiple stages or continuous-contacting devices may be useful. The principles of such operations are analogous to those of other countercurrent contacting operations—for example, distillation, absorption, and extraction.

Variables other than temperature and the presence of other solutes can influence solubility. For example, the effect of a nonsolvent on solubility sometimes is used to bring about recovery of a solute. Figure 5 shows the solubility of L-serine in aqueous solutions containing varying amounts of methanol. Note that increasing methanol content reduces the solubility by more than an order of magnitude, and this characteristic can be used to obtain a high yield in the recovery of L-serine.

There is increasing interest in the crystallization of solutes from supercritical-fluid solvents. In such instances, solubilities often are correlated by an equation of state. Such concepts are beyond the scope of the current discussion but are presented elsewhere in the encyclopedia.

Although this discussion provides insight to the types of solubility behavior that can be exhibited by various systems, it is by no means a complete survey of the topic. Extensive solubility data and descriptions of more complex equilibrium behavior can be found in the literature. Published data usually consist of the influence of temperature on the solubility of a pure solute in a pure solvent; seldom are effects of other solutes, co-solvents, or pH considered. As a consequence, solubility data on a system of interest should be measured experimentally, and the solutions used in the experiments should be as similar as possible to those expected in the process. Even if a crystallizer has been designed and the process is operational, obtaining solubility data using mother liquor drawn from the crystallizer or a product stream would be wise. Moreover, the solubility should be checked periodically to see if it has changed due to changes in the upstream operations or raw materials.

There have been advances in the techniques by which solid-liquid equilibria can be correlated and, in some cases, predicted. These are described in references on phase-equilibrium thermodynamics.

B. Mass and Energy Balances

Illustrating the formulation of mass and energy balances is simplified by restricting the analysis to systems whose crystal growth kinetics are sufficiently fast to utilize essentially all of the supersaturation provided by the crystallizer; in other words, the product solution

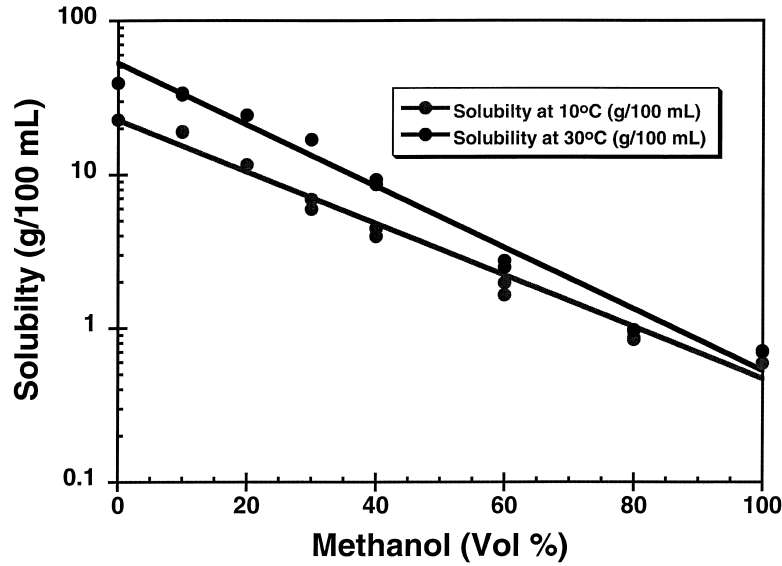


FIGURE 5 Effect of methanol on solubility of L-serine.

is assumed to be saturated. Under such conditions (referred to in the crystallization literature as Class II or fast-growth behavior), the solute concentration in the mother liquor can be assigned a value corresponding to saturation. Should the supersaturation in the mother liquor be so great as to affect the solute balance, the operation is said to follow Class I or slow-growth behavior. In Class I behavior, the operating conditions affect the rate at which solute is crystallized, and an expression coupling the rate of growth to a solute balance must be used to describe the system. Such treatment will be considered beyond the scope of this discussion.

The solution of mass and energy balances requires solubility and enthalpy data on the system of interest. Various methods of presenting solubility data were given earlier, and the use of solubilities to estimate crystal production rates from a cooling crystallizer was demonstrated by the discussion of Eq. (2). Subsequent to determining the yield, the rate at which heat must be removed from the crystallizer can be calculated from an energy balance:

$$m_C \hat{H}_C + m_L \hat{H}_L - m_F \hat{H}_F = Q \quad (3)$$

where m_F , m_C , and m_L are feed rate, crystal production rate, and mother liquor flow rate, respectively; $\hat{H}$ is specific enthalpy of the stream corresponding to the subscript; and Q is the required rate of heat transfer to the crystallizer. As m_F , m_C , and m_L are known or can be calculated from a simple mass balance, determination of Q requires estimation of specific enthalpies. These are most conveniently obtained from enthalpy-composition diagrams, which are available in the general literature for a number of substances.

If specific enthalpies are unavailable, they can be estimated based on defined reference states for both solute and solvent. Often the most convenient reference states are crystalline solute and pure solvent at an arbitrarily chosen reference temperature. The reference temperature selected usually corresponds to that at which the heat of crystallization, $\Delta \hat{H}_c$, of the solute is known. (The heat of crystallization is approximately equal to the negative of the heat of solution.) For example, if the heat of crystallization is known at T_{ref} , then reasonable reference conditions would be the solute as a solid and the solvent as a liquid, both at T_{ref} . The specific enthalpies could be estimated then as:

$$\hat{H}_F = x_F \Delta \hat{H}_c + C_{pF}(T - T_{\text{ref}}) \quad (4)$$

$$\hat{H}_C = C_{pC}(T - T_{\text{ref}}) \quad (5)$$

$$\hat{H}_L = x_L \Delta \hat{H}_c + C_{pL}(T - T_{\text{ref}}) \quad (6)$$

where x_F and x_L are the mass fractions of solute in the feed and mother liquor, respectively. All that is required now to determine the required rate of heat transfer is the indicated heat capacities, which can be estimated based on system composition or measured experimentally.

Now suppose some of the solvent is evaporated in the crystallizer. Independent balances can be written on total and solute masses:

$$m_F = m_V + m_L + m_C \quad (7)$$

$$x_F m_F = x_L m_L + x_C m_C \quad (8)$$

Assuming that the streams leaving the crystallizer are in equilibrium, there is a relationship between the temperature (or pressure) at which the operation is conducted

and x_L and x_C . In addition, an energy balance must be satisfied:

$$m_F \hat{H}_F + Q = m_V \hat{H}_V + m_L \hat{H}_L + m_C \hat{H}_C \quad (9)$$

The specific enthalpies in the above equation can be determined as described earlier, provided the temperatures of the product streams are known. Evaporative cooling crystallizers (described more completely in Section V) operate at reduced pressure and may be considered adiabatic. In such circumstances, Eq. (9) is modified by setting $Q = 0$. As with many problems involving equilibrium relationships and mass and energy balances, trial-and-error computations are often involved in solving Eqs. (7) through (9).

III. NUCLEATION AND GROWTH KINETICS

The kinetics of crystallization have constituent phenomena in crystal nucleation and growth. The rates at which these occur are dependent on driving forces (usually expressed as supersaturation), physical properties, and process variables, but relationships between these quantities and crystallization kinetics often are difficult to express quantitatively. As a result, empirical or qualitative links between a process variable and crystallization kinetics are useful in providing guidance in crystallizer design and operation and in developing strategies for altering the properties of crystalline products.

Nucleation and growth can occur simultaneously in a supersaturated environment, and the relative rates at which these occur are primary determinants of the characteristics of the crystal size distribution; one way of influencing product size distributions is through the control of variables such as supersaturation, temperature, and mixing characteristics. Obviously, those factors that increase nucleation rates relative to growth rates lead to a crystal size distribution consisting of smaller crystals. In the discussion that follows, an emphasis will be given to the general effects of process variables on nucleation and growth, but the present understanding of these phenomena does not allow quantitative *a priori* prediction of the rates at which they occur.

A. Supersaturation

Supersaturation is the thermodynamic driving force for both crystal nucleation and growth; and therefore, it is the key variable in setting the mechanisms and rates by which these processes occur. It is defined rigorously as the deviation of the system from thermodynamic equilibrium and is quantified in terms of chemical potential,

$$\Delta\mu_i = \mu_i - \mu_i^* = RT \ln \frac{a_i}{a_i^*} \quad (10)$$

where μ_i is the chemical potential of solute i at the existing conditions of the system, μ_i^* is the chemical potential of the solute equilibrated at the system conditions, and a_i and a_i^* are activities of the solute at the system conditions and at equilibrium, respectively. Less abstract definitions involving measurable system quantities are often used to approximate supersaturation; these involve either temperature or concentration (mass or moles of solute per unit volume or mass of solution or solvent) or mass or mole fraction of solute. Recommendations have been made that it is best to express concentration in terms of moles of solute per unit mass of solvent. For systems that form hydrates, the solute should include the water of hydration, and that water should be deducted from the mass of solvent.

Consider, for example, a system at temperature T with a solute concentration C , and define the equilibrium temperature of a solution having a concentration C as T^* and the equilibrium concentration of a solution at T as C^* . These quantities may be used to define the following approximate expressions of supersaturation:

1. The difference between the solute concentration and the concentration at equilibrium, $\Delta C_i = C_i - C_i^*$
2. For a solute whose solubility in a solvent increases with temperature, the difference between the temperature at equilibrium and the system temperature, $\Delta T = T^* - T$
3. the supersaturation ratio, which is the ratio of the solute concentration and the equilibrium concentration, $S_i = C_i / C_i^*$
4. The ratio of the difference between the solute concentration and the equilibrium concentration to the equilibrium concentration, $\sigma_i = (C_i - C_i^*) / C_i^* = S_i - 1$, which is known as relative supersaturation.

Any of the above definitions of supersaturation can be used over a moderate range of system conditions, but as outlined in the following paragraph, the only rigorous expression is given by Eq. (10).

The definitions of supersaturation ratio and relative supersaturation can be extended to any of the other variables used in the definition of supersaturation. For example, defining $S_{a_i} = a_i / a_i^*$ gives:

$$\frac{\Delta\mu_i}{RT} = \ln S_{a_i} = \ln \frac{\gamma_i C_i}{\gamma_i^* C_i^*} \quad (11)$$

Therefore, for ideal solutions or for $\gamma_i \approx \gamma_i^*$,

$$\frac{\Delta\mu_i}{RT} \approx \ln \frac{C_i}{C_i^*} = \ln S_i \quad (12)$$

Furthermore, for low supersaturations (say, $S_i < 1.1$),

$$\frac{\Delta\mu_i}{RT} \approx S_i - 1 = \sigma_i \quad (13)$$

The simplicity of Eq. (13) results in the use of relative supersaturation in most empirical expressions for nucleation and growth kinetics. While beguilingly simple, and correct in limiting cases, great care should be taken in extending such expressions beyond conditions for which the correlations were developed.

For ionic solutes, $a_i = a_{\pm}^{\nu}$, which leads to $S_{a_i} = (a_{\pm}/a_{\pm}^*)^{\nu}$ and

$$\frac{\Delta\mu_i}{RT} = \nu \ln S_{a_i} = \nu \ln \frac{\gamma_{i\pm} C_i}{\gamma_{i\pm}^* C_i^*} \quad (14)$$

Again, for $\gamma_{i\pm} \approx \gamma_{i\pm}^*$,

$$\frac{\Delta\mu_i}{RT} \approx \nu \ln \frac{C_i}{C_i^*} = \nu \ln S_i \quad (15)$$

B. Primary Nucleation

The term *primary nucleation* is used to describe both homogeneous and heterogeneous nucleation mechanisms in which solute crystals play no role in the formation of new crystals. Primary nucleation mechanisms involve the formation of crystals through a process in which constituent crystal units are stochastically combined. Both homogeneous and heterogeneous nucleation require relatively high supersaturations, and they exhibit a high-order dependence on supersaturation. As will be shown shortly, the high-order dependence has a profound influence on the character of crystallization processes in which primary nucleation is the dominant means of crystal formation.

The classical theoretical treatment of primary nucleation that produces a spherical nucleus results in the expression:

$$\begin{aligned} B^{\circ} &= A \exp\left(-\frac{16\pi\epsilon_{\text{surf}}^3 v^2}{3k^3 T^3 [\ln(\sigma + 1)]^2}\right) \\ &\approx^{\sigma < 0.1} A \exp\left(-\frac{16\pi\epsilon_{\text{surf}}^3 v^2}{3k^3 T^3 \sigma^2}\right) \end{aligned} \quad (16)$$

where k is the Boltzmann constant, ϵ_{surf} is the interfacial surface energy per unit area, v is molar volume of the crystallized solute, and A is a constant.

The theory shows that the most important variables affecting the rates at which primary nucleation occur are interfacial energy ϵ_{surf} , temperature T , and supersaturation σ . The high-order dependence of nucleation rate on these three variables, especially supersaturation, is important because, as shown by an examination of Eq. (16), a small change in any of the three variables could produce an enormous change in nucleation rate. Such behavior gives rise to the often observed phenomenon of having a clear

liquor transformed to a slurry of very fine crystals with only a slight increase in supersaturation, for example by decreasing the solution temperature.

The effect of exogenous solid matter (as in heterogeneous nucleation) in the supersaturated solution is equivalent to that of a catalyst in a reactive mixture. Namely, it is to reduce the energy barrier to the formation of a new phase. In effect, the solid matter reduces the interfacial energy ϵ_{surf} by what may amount to several orders of magnitude.

The classical nucleation theory embodied in Eq. (16) has a number of assumptions and physical properties that cannot be estimated accurately. Accordingly, empirical power-law relationships involving the concept of a metastable limit have been used to model primary nucleation kinetics:

$$B^{\circ} = k_N \sigma_{\text{max}}^n \quad (17)$$

where k_N and n are parameters fit to data and σ_{max} is the supersaturation at which nuclei are observed when the system is subjected to a specific protocol. Although Eq. (17) is based on empiricism, it is consistent with the more fundamental Eq. (16).

C. Secondary Nucleation

Secondary nucleation is the formation of new crystals through mechanisms involving existing solute crystals; in other words, crystals of the solute *must* be present for secondary nucleation to occur. Several features of secondary nucleation make it important in the operation of industrial crystallizers: First, continuous crystallizers and seeded batch crystallizers have crystals in the magma that can participate in secondary nucleation mechanisms. Second, the requirements for the mechanisms of secondary nucleation to be operative are fulfilled easily in most industrial crystallizers. Finally, many crystallizers are operated in a low supersaturation regime so as to maximize yield, and at such supersaturations the growth of crystals is more likely to produce desired morphologies and high purity; these low supersaturations can support secondary nucleation but not primary nucleation.

1. Mechanisms

Secondary nucleation can occur through several mechanisms, including initial breeding, contact nucleation (also known as collision breeding), and shear breeding. Although a universal expression for the kinetics of secondary nucleation does not exist, a working relationship often can be obtained by correlating operating data from a crystallizer with a semi-empirical expression. Guidance as to the

form of the expression and the variables that it should include can be obtained by understanding the various mechanisms of secondary nucleation.

Initial breeding results from immersion of seed crystals in a supersaturated solution, and it is thought to be caused by dislodging extremely small crystals that were formed on the surface of larger crystals during drying. Although this mechanism is unimportant in continuous and unseeded batch crystallization, it can have a significant impact on the operation of seeded batch crystallizers. The number of crystals formed by initial breeding, has been found to be proportional to the surface area of crystals used to seed a batch crystallizer. Characteristics of the resulting distribution are affected strongly by the growth kinetics of nuclei resulting from initial breeding, and the phenomenon of growth-rate dispersion (which will be discussed later) can lead to erroneous conclusions regarding the nucleation kinetics.

Shear breeding results when supersaturated solution flows by a crystal surface and carries with it crystal precursors believed formed in the region of the growing crystal surface. High supersaturation is required for shear breeding to produce significant numbers of nuclei.

Contact nucleation in industrial processes results from collisions of crystals with the impeller used for circulation of the magma or with other crystallizer internals such as baffles, pipe and crystallizer walls, and even other crystals. Careful experimental studies have shown that the number of crystals produced by collisions between crystals and these objects depends upon the collision energy, supersaturation at impact, supersaturation at which crystals mature, material of the impacting object, area and angle of impact, and system temperature. The collision energy for contact nucleation is small and does not necessarily result in the macroscopic degradation or attrition of the contacted crystal.

Nucleation from collisions between crystals in the circulating magma and the rotor in a circulation pump or an agitator usually dominate nucleation resulting from other collisions. The operating variables in systems of this type can be manipulated to some extent, thereby modifying nucleation rates and the concomitant crystal size distribution. For example, internal classification can be used to keep larger crystals away from energetic collisions with an impeller, but doing so may create other problems with stability of the crystal size distribution. The rotational speed of an impeller can be changed if there are appropriate controls on the pump or agitator. Caution must be exercised, however, for a reduction in circulation velocity can reduce heat-transfer coefficients and increase fouling (encrustation) on heat-transfer surfaces. Moreover, the crystals in the magma must be kept suspended or crystal morphology and growth rates could be affected adversely. Impact

energy may have a high-order dependence on rotational speed and, if that is the case, modest changes in this variable could alter nucleation rates substantially. The fraction of the impact energy transmitted from an impeller to the crystal can be manipulated by changing the material of construction of the impeller. The influence of using soft materials to coat impellers or crystallizer internals may vary from one crystalline system to another; those systems in which the crystal face is soft may be more susceptible to nucleation rate changes than those crystalline systems where the face is hard.

Supersaturation has been observed to affect contact nucleation, but the mechanism by which this occurs is not clear. There are data that infer a direct relationship between contact nucleation and crystal growth; these data showed that the number of nuclei produced by an impact was proportional to the linear growth rate of the impacted face. This could indicate that the effect of supersaturation is to alter growth rates and, concomitantly, the characteristics of the impacted crystal faces; alternatively, what appears to be a mechanistic relationship actually could be a result of both nucleation and growth depending upon supersaturation.

Another theory that could account for the effect of supersaturation on contact nucleation is based on the view that nuclei formed cover a range of sizes that includes the critical nucleus. Since only the nuclei larger than the critical nucleus are stable, the relationship of the size of the critical nucleus to supersaturation reflects the dependence of contact nucleation on supersaturation. This concept, which has been referred to as a survival theory, seems to have been refuted by measurements of the sizes of crystals formed by collisions. These sizes are much larger than the critical nucleus, and the survival theory would have little influence on the number of nuclei that survive.

Evidence of the formation of polymolecular clusters in supersaturated solutions may provide a mechanistic interpretation of the effect of supersaturation on contact nucleation kinetics. These clusters may participate in nucleation, although the mechanism by which this would occur is not clear. One model that has been proposed, however, calls for the formation of a semi-ordered region consisting of molecular clusters awaiting incorporation into the crystal lattice. Collisions or fluid shear of the region containing high cluster concentrations could then result in these clusters serving as secondary nuclei. In such a model, the variables that influence formation and diffusion of the clusters also influence crystal growth rates and nucleation.

2. Kinetic Expressions

Irrespective of the actual mechanisms by which contact nucleation occurs, empirical power-law expressions

provide a useful means of correlating nucleation kinetics and using the resulting correlations in process analysis and control. The correlations generally take the form:

$$B^\circ = k_N \sigma^i M_T^j N^k \quad (18)$$

where k_N , i , j , k are positive parameters obtained from data correlation, M_T is the magma density (mass of solids per unit volume of slurry or solvent in the magma), and N is the rotational velocity of the impeller or pump rotor. For convenience, either crystal growth rate or mean residence time, both of which are directly related to supersaturation, may be substituted for σ in Eq. (18).

If primary nucleation dominates the process, i tends to larger values (say greater than 3), j and k approach zero, and Eq. (18) approaches Eq. (17). Should crystal-impeller and/or crystal-crystallizer impacts dominate, j approaches 1; on the other hand, if crystal-crystal contacts dominate, j approaches 2.

The ease with which nuclei can be produced by contact nucleation is a clear indication that this mechanism is dominant in many industrial crystallization operations. Research on this nucleation mechanism is continuing with the objective of building an understanding of the phenomenon that will allow its successful inclusion in models describing commercial systems.

D. Fundamentals of Crystal Growth

Crystal growth rates may be expressed as (1) the linear advance rate of an individual crystal face, (2) the change in a characteristic dimension of a crystal, or (3) the rate of change in mass of a crystal or population of crystals. These different expressions are related through crystal geometry; it is often convenient to use the method of measurement as the basis of the growth rate expression or, in certain instances, the method used to analyze a crystallization process will require that growth rate be defined in a specific way. For example, the use of a population balance to describe crystal size distribution requires that growth rate be defined as the rate of change of a characteristic dimension.

Single-crystal growth kinetics involve the advance rate of an individual crystal face normal to itself or the rate of change in crystal size associated with exposure to a super-saturated solution. The advance rate of a single crystal face can be quantified by observation of the face through a calibrated eyepiece of an optical microscope, which allows examination of the structure of the advancing crystal face and isolation of surface-reaction kinetics from mass-transfer kinetics (these phenomena will be discussed later). An additional advantage of single-crystal systems is that it is possible to examine crystal growth kinetics without interference from competing processes such as nucleation.

Multicrystal-magma studies usually involve examination of the rate of change of a characteristic crystal dimension or the rate of increase in the mass of crystals in a magma. The characteristic dimension in such analyses depends upon the method used in the determination of crystal size; for example, the second largest dimension is measured by sieve analyses, while an equivalent spherical diameter is determined by both electronic zone sensing and laser light scattering instruments. A relationship between these two measured dimensions and between the measured quantities and the actual crystal dimensions can be derived from appropriate shape factors. Volume and area shape factors are defined by the equations:

$$v_{\text{crys}} = k_{\text{vol}} L^3 \quad \text{and} \quad a_{\text{crys}} = k_{\text{area}} L^2 \quad (19)$$

where v_{crys} and a_{crys} are volume and area of a crystal, k_{vol} and k_{area} are volume and area shape factors, and L is the characteristic dimension of the crystal. Suppose an equivalent spherical diameter L_{sphere} is obtained from an electronic zone-sensing instrument, and the actual dimensions of the crystal are to be calculated. Assume for the sake of this example that the crystals have a cubic shape. Let L_{cube} be the edge length of the crystal and $k_{\text{vol}}^{\text{sphere}}$ and $k_{\text{vol}}^{\text{cube}}$ be the volume shape factors for a sphere and a cube, respectively. Since the volume of the crystal is the same, regardless of the arbitrarily defined characteristic dimension,

$$v_{\text{crys}} = k_{\text{vol}}^{\text{sphere}} L_{\text{sphere}}^3 = k_{\text{vol}}^{\text{cube}} L_{\text{cube}}^3 \quad (20)$$

Since $k_{\text{vol}}^{\text{sphere}}$ is $\pi/6$ and $k_{\text{vol}}^{\text{cube}}$ is 1.0, the numerical relationship between L_{cube} and L_{sphere} is given by:

$$L_{\text{cube}} = \left(\frac{k_{\text{vol}}^{\text{sphere}}}{k_{\text{vol}}^{\text{cube}}} \right)^{1/3} L_{\text{sphere}} = \left(\frac{\pi}{6} \right)^{1/3} L_{\text{sphere}} \quad (21)$$

The rate of change of a crystal mass dm_{crys}/dt can be related to the rate of change in the crystal characteristic dimension ($dL/dt = G$) by the equation:

$$\frac{dm_{\text{crys}}}{dt} = \frac{d(\rho k_{\text{vol}} L^3)}{dt} = 3\rho k_{\text{vol}} L^2 \left(\frac{dL}{dt} \right) \quad (22)$$

where ρ is crystal density. Since $k_{\text{area}} = a_{\text{crys}}/L^2$,

$$\frac{dm_{\text{crys}}}{dt} = 3\rho (k_{\text{vol}}/k_{\text{area}}) a_{\text{crys}} G \quad (23)$$

At least two resistances contribute to the kinetics of crystal growth. These resistances apply to (1) integration of the crystalline unit (e.g., solute molecules) into the crystal surface (i.e., lattice), and (2) molecular diffusion or bulk transport of the unit from the surrounding solution to the crystal surface. As aspects of molecular diffusion and mass transfer are covered elsewhere, the current discussion will focus only on surface incorporation.

1. Mechanisms

Among the many models that have been proposed to describe surface-reaction kinetics are those that assume crystals grow by layers and others that consider growth to occur by the movement of a continuous step. Each physical model results in a specific relationship between growth rate and supersaturation and, although none can predict growth kinetics *a priori*, insights regarding the effects of process variables on growth can be obtained. Because of the extensive literature on the subject, only the key aspects of the physical models and (in one case) the resulting relationship between growth and supersaturation predicted by each theory will be discussed here.

The model used to describe the growth of crystals by layers is based on a two-step, birth-and-spread mechanism. In one of the steps (birth) a two-dimensional nucleus is formed on the crystal surface, and in the second step (spread) the two-dimensional nucleus grows to cover the crystal surface. When one or the other of the steps is controlling growth rates, simplifications of the more complicated dependence of growth rate on supersaturation can be developed to give what are known as the mononuclear two-dimensional nucleation theory and the polynuclear two-dimensional nucleation theory. In the mononuclear two-dimensional nucleation theory, surface nucleation occurs at a finite rate while the spreading across the surface occurs at an infinite rate. The reverse is true for the polynuclear two-dimensional nucleation theory. Theoretical relationships have been derived between growth rate and supersaturation for each of these conditions but are considered beyond the scope of this discussion.

The screw-dislocation theory (sometimes referred to as the BCF theory because of its development by Burton, Cabrera, and Frank) is based on a mechanism of continuous movement in a spiral or screw of a step or ledge on the crystal surface. The theory shows that the dependence of growth rate on supersaturation can vary from a parabolic relationship at low supersaturations to a linear relationship at high supersaturations. In the BCF theory, growth rate is given by:

$$G = k_G \left(\frac{\epsilon \sigma^2}{b} \right) \tanh \left(\frac{b}{\epsilon \sigma} \right) \quad (24)$$

where ϵ is screw dislocation activity and b is a system-dependent quantity that is inversely proportional to temperature. It can be shown that the dependence of growth rate on supersaturation is linear if the ratio $b/\epsilon\sigma$ is large, but the dependence becomes parabolic as the ratio becomes small. It is possible, then, to observe variations in the dependence of growth rate on supersaturation for a given crystal-solvent system.

An empirical approach also can be used to relate growth kinetics to supersaturation by simply fitting growth-rate data with a power-law function of the form:

$$G = k_G \sigma^g \quad (25)$$

where k_G and g are system-dependent constants. Such an approach is valid over modest ranges of supersaturation, and the power-law function approximates the fundamental expressions derived from the above models.

2. Impurities

The presence of impurities can alter growth rates substantially, usually by decreasing them. Furthermore, as described in Section IV.B, impurities can alter crystal morphology through their effects on the growth rates of crystal faces. Mechanisms include: (1) adsorption of an impurity on the crystal surface or at specific growth sites such as kinks, thereby blocking access to the site by a growth unit; (2) formation of complexes between an impurity and a growth unit; and (3) incorporation of an impurity into a growing crystal and creating defects or repelling the addition of a growth unit to the subsequent crystal layer. Few of these mechanistic views result in predictive capabilities, and it is usual to rely on experimental data that are often correlated empirically.

Because impurities most often result in reduced crystal growth rate, feedstocks to laboratory and bench-scale units should be as similar as possible to that expected in the full-scale unit. The generation of impurities in upstream process units can depend on the way those units are operated, and protocols of such units should follow a consistent practice. It is equally important to monitor the composition of recycle streams so as to detect any accumulation of impurities that might lead to a reduction in growth rates.

The solvent from which a material is crystallized influences crystal morphology and growth rate. These effects have been attributed to two sets of factors. One has to do with the effects of solvent on viscosity, density, and diffusivity and, therefore, mass transfer. The second factor is concerned with the structure of the interface between crystal and solvent; a solute-solvent system that has a high solubility is likely to produce a rough interface and, concomitantly, large crystal growth rates.

E. Crystal Growth in Mixed Crystallizers

Population balances on crystals in a crystallizer require a definition of growth rates in terms of the rate of change of a characteristic dimension:

$$G = \frac{dL}{dt} \quad (26)$$

Furthermore, the solution of a differential population balance requires that the relationship between growth rate and size of the growing crystals be known. When all crystals in the magma grow at a constant and identical rate, the crystal-solvent system is said to follow the McCabe ΔL law, while systems that do not are said to exhibit anomalous growth.

Two theories have been used to explain growth-rate anomalies: size-dependent growth and growth-rate dispersion. As with systems that follow the ΔL law, anomalous growth by crystals in a multicrystal magma produces crystal populations with characteristic forms. Unfortunately, it is difficult to determine the growth mechanism from an analysis of these forms. This means that either size-dependent growth or growth-rate dispersion may be used to correlate population density data without a certainty that the correct source of anomalous growth has been identified. Determining the actual source of anomalous growth is not trivial, but it may be worthwhile since alignment between a mathematical model and system behavior enhances the utility of the model.

Size-dependent crystal growth results when the rate of growth depends on the size of the growing crystal. Certainly, this may be the case if bulk transport is the controlling resistance to crystal growth, and the literature abounds with expressions for the appropriate mass-transfer coefficients. In the more common situation in which surface integration controls growth rate, there are no mechanistic relationships between growth rate and crystal size, and simple empirical expressions are called upon for that purpose.

Growth-rate dispersion is the term used to describe the behavior of similar sized crystals in the same population exhibiting different growth rates or growth rates that vary with time. The consequences of growth-rate dispersion are illustrated in Fig. 6, which shows the growth of a crystal population that has been immersed in a supersaturated solution. The spread of the distribution increases as the crystal population grows; the slower growing crystals form the tail of the advancing distribution while the faster growing ones form the leading edge. If all crystals in the population grew at the same rate, the distribution would advance uniformly along the size axis. Two causes of growth-rate dispersion have been observed. In one, the growth rate of each crystal in a population is nearly constant, but crystals in the population may grow at a different rate; in the other, the growth rate of an individual crystal fluctuates about a mean value.

The consequences of anomalous growth depends upon the process involved, and this will be pointed out in the discussion on population balances.

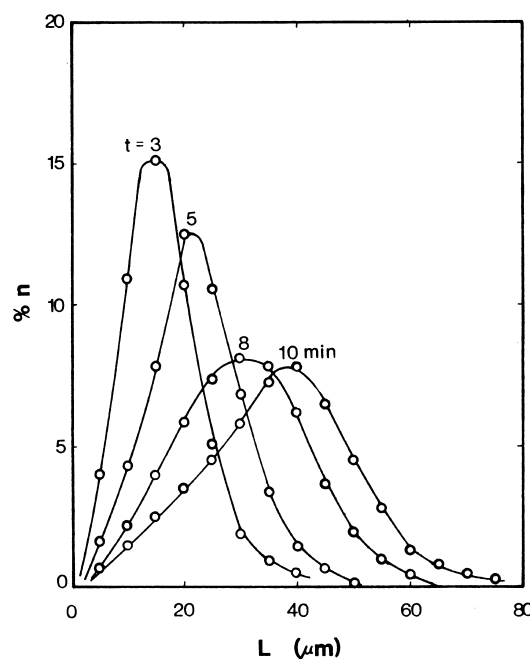


FIGURE 6 Transient population density plot showing growth-rate dispersion.

IV. PURITY, MORPHOLOGY, AND SIZE DISTRIBUTIONS

Crystal properties can be divided into two categories: those based on the individual crystal and those involving all crystals of a given population. The three characteristics of the section title compose what is often called *crystal quality*. They represent, along with yield, the most important criteria in the operation of a crystallizer. In the following discussion, some of the factors that influence purity and morphology are described and an introduction is given to methods of quantifying crystal size distributions.

A. Crystal Purity

The purity of a crystalline product depends on the nature of the other species in the mother liquor from which the crystals are produced, the physical properties of the mother liquor, and the processing that occurs between crystallization and the final product (downstream processing). Impurities can find their way into the final product through a number of mechanisms: the formation of occlusions, trapping of mother liquor in physical imperfections of the crystals or agglomerates, adsorption of species onto crystal surfaces, as part of chemical complexes (hydrates or solvates), or through lattice substitution.

Occlusions find their way into the crystal structure when the supersaturation in close proximity to the crystal surface is high enough to lead to an unstable surface. Such instability leads to the creation of dendrites, which then join to trap mother liquor in pools of liquid within the crystal. Occlusions are often visible and can be avoided through careful control of the supersaturation in the crystallizer.

Mother liquor can be flushed from a cake of crystals on a filter or centrifuge by washing with a liquid that also may dissolve a small portion of the cake mass. To be effective, the wash liquid must be spread uniformly over the cake and flow through the porous material without significant channeling. Such washing is hindered when the crystals themselves have significant cracks, crevices, or other manifestations of breakage or the mother liquor has a viscosity that is significantly greater than the wash liquid. In the latter event, significant channeling (also called *fingering*) may reduce the effectiveness of the wash process.

Lattice substitution requires that the incorporated impurity be of similar size and function to the primary crystallizing species. In other words, the impurity must fit into the lattice without causing significant dislocations. An example of such a system is found in the crystallization of L-isoleucine in the presence of trace quantities of L-leucine. The two species have similar molecular structures, differing only by one carbon atom in the position of a methyl side group. In this system, the incorporation of L-leucine in L-isoleucine crystals is proportional to the concentration of L-leucine in the mother liquor. Moreover, the shape of the recovered crystals changes as the content of L-leucine in recovered crystal increases.

B. Crystal Morphology

Both molecular and macroscopic concepts are important in crystal morphology. Molecular structures (i.e., the arrangements of molecules in specific lattices) can greatly influence the properties of a crystalline species and variations from a single structure lead to the prospect of polymorphic systems. In such systems, the molecular species of the crystal can occupy different locations depending on the conditions at which the crystal is formed, and both microscopic and macroscopic properties of the crystal can vary depending on the polymorph formed. There is, in general, a single stable polymorph for prevailing conditions, but that polymorph may not have been formed during the crystallization process. In such cases, system thermodynamics will tend to force transformation from the unstable polymorph to the stable one at rates that may vary from being nearly instantaneous to infinitely slow. Additional discussion of the molecular structures of crystalline materials has been provided elsewhere.

The characteristic macroscopic shape of a crystal results, in large measure, from the internal lattice structure; surfaces are parallel to planes formed by the constituent units of the crystal. Moreover, although the Law of Constant Interfacial Angles is a recognition that angles between corresponding faces of all crystals of a given substance are constant, the faces of individual crystals of that substance may exhibit varying degrees of development. As a result, the general shape or habit of a crystal may vary considerably.

Crystal morphology (i.e., both form and shape) affects crystal appearance; solid-liquid separations such as filtration and centrifugation; product-handling characteristics such as dust formation, agglomeration, breakage, and washing; and product properties such as bulk density, dissolution kinetics, catalytic activity, dispersability, and caking.

The shape of a crystal can vary because the relative rates of growth of crystal faces can change with system conditions; faster growing faces become smaller than faces that grow more slowly and in the extreme may disappear from the crystal altogether. For illustration, consider the two-dimensional crystal shown in Fig. 7a and the process variables that would cause the habit to be modified to the forms shown in Figs. 7b and c. The shape of the crystal depends on the ratio of the growth rate of the horizontal faces, G_h , to the growth rate of the vertical faces, G_v . For the shapes shown in Fig. 7,

$$\left(\frac{G_h}{G_v}\right)_b < \left(\frac{G_h}{G_v}\right)_a < \left(\frac{G_h}{G_v}\right)_c \quad (27)$$

Growth rates depend on the presence of impurities, system temperature, solvent, mixing, and supersaturation, and the importance of each may vary from one crystal face to another. Consequently, an alteration in any or all of these variables can result in a change of the crystal shape.

Modeling intermolecular and intramolecular interactions through molecular mechanics calculations has advanced significantly in the past decade, and it has provided the basis for prediction of the equilibrium shape of

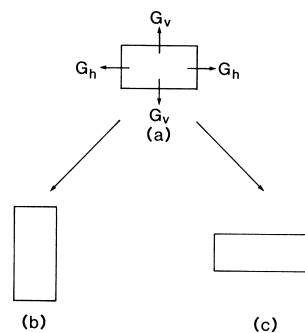


FIGURE 7 Effect of facial growth rates on crystal shape.

a known crystalline species. While not yet uniformly successful on a quantitative basis, the definition and modeling of crystal lattice potential energy equations has provided an understanding of crystal growth and morphology on the molecular level. Derivation of external crystal morphology from internal lattice structures via simulation has been proven possible for several organic compounds. Numerical minimization techniques, coupled with the appropriate valence and nonbonded energy expressions, have enabled accurate determination of favorable molecular arrangements within a wide variety of molecular crystals.

The shape of crystals obtained as a result of following a specific crystallization protocol may be unsatisfactory and, as a result, methods for modifying the habit of considerable interest. The predictive capabilities cited in the preceding paragraph are of great utility in such an instance as they may be used to determine factors leading to the unsatisfactory shape and guide subsequent experiments in which a more desirable shape is sought. Inevitably, such a search involves extensive laboratory or bench-scale experiments to determine processing variations that will lead to a desired crystal shape.

As an example of the variations in shape that can be exhibited by a single crystalline material, consider the forms of potassium sulfate shown in Fig. 8. Clearly, the processing characteristics and particulate properties of the differently shaped potassium sulfate crystals will vary.

The mechanisms and variables affecting crystal shape can be categorized as follows:

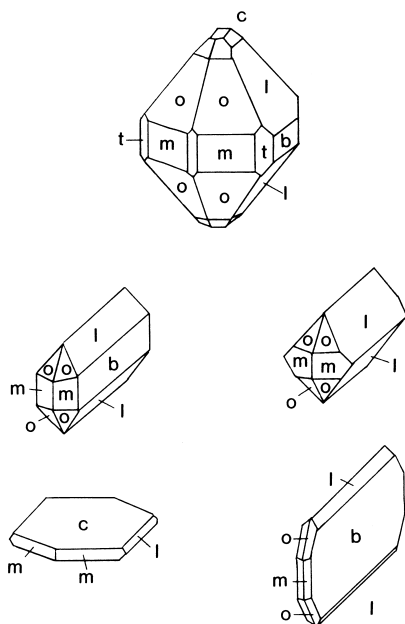


FIGURE 8 Shapes of K_2SO_4 crystals. [From Mullin, J. W. (1993). "Crystallization," 3rd ed. Butterworth-Heinemann, London. With permission.]

1. Intrinsic growth rates

- a. *Temperature*: The growth rates of individual crystal faces depend on temperature, typically following an Arrhenius rate law:

$$G = G_0 \exp\left(-\frac{\Delta E_G}{RT}\right) \quad (28)$$

If different crystal faces have different activation energies, variation of the temperature at which crystallization takes place modifies individual growth rates to varying degrees and results in a modified crystal shape.

- b. *Mixing*: The intensity of mixing may determine the degree to which bulk mass transfer is involved in growth kinetics, and this can influence the resulting crystal shape.
- c. *Supersaturation*: The dependence of growth kinetics on supersaturation may vary from one crystal face to another. Accordingly, different prevailing supersaturations can lead to different crystal shapes.

2. Interfacial behavior

- a. *Solvent*: Different solvents exhibit different interactions with crystal faces and can alter crystal shape. A change in solvent also can alter the stoichiometry of the crystal (e.g., from a hydrated to an anhydrate stoichiometry), which can produce crystals with quite different morphology.
- b. *Surfactants*: Addition of a surfactant to a crystallizing system can influence the crystal shape in a manner illustrated schematically in Fig. 9. Here, surfactant molecules are shown being attracted to crystal faces in varying ways; the hydrophilic head groups favor the horizontal faces, while the hydrophobic tail groups are preferentially attracted to the vertical faces. A growth unit must displace the surfactant to approach a growing crystal face. As hydrophilic interactions are typically much stronger than hydrophobic ones, the growth unit preferentially enters the vertical faces and growth in the horizontal direction is favored.

3. Access to growth site

- a. *Blockage by species attracted to growth site*: Impurities may preferentially locate at a kink or other favored growth site and block growth at that site. A difference in the character of the kink or growth site from one face to another could result in modification of the crystal shape.
- b. *Species partially fitting into crystal lattice*: In these instances, an impurity molecule is comprised of two parts, one that fits into the crystal lattice

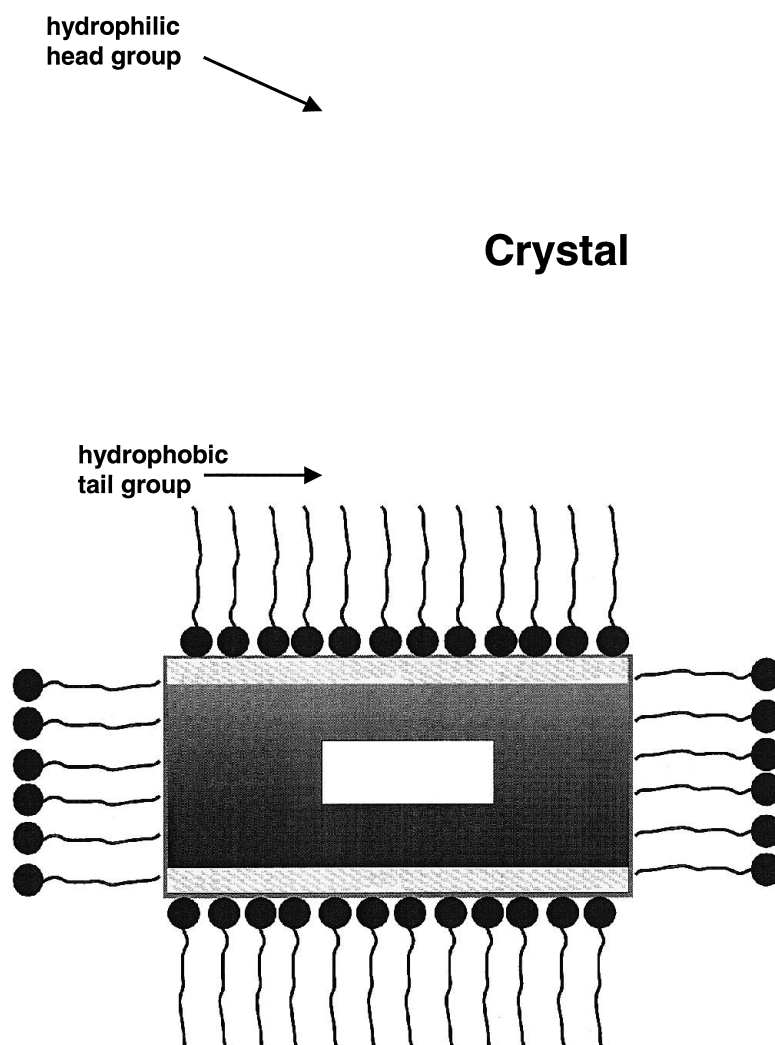


FIGURE 9 Attachment of surfactant molecules to crystal surfaces.

and a second that does not. The part that does not fit repulses incoming growth units or causes significant interatomic stress because of its position in the crystal lattice. If such species are purposely added to the crystallizing system to modify crystal morphology, they are referred to as tailor-made additives.

C. Crystal Size Distributions

Most crystallization processes produce particles whose sizes cover a range of varying breadth. If the particles consist of single crystals, the resulting distribution is a crystal size distribution (CSD); on the other hand, if the particles consist of agglomerates or other combination of multiple crystals, the distribution is a particle size distribution. In either case, the distribution is expressed in terms of either population (number) or mass. The popu-

lation distribution relates the number of crystals at each size to the size, while the mass distribution expresses how mass is distributed over the size range. In the following paragraphs, methods for describing and using distribution functions will be outlined.

Size distribution is a major determinant of the properties of crystalline products, especially appearance, and to downstream processing and handling of crystalline materials. Solid-liquid separation by filtration or centrifugation can be straightforward with a desired CSD, but it can be disastrous when an inappropriate one increases resistance to liquor flow through a filter or centrifuge cake. Likewise, CSD affects other downstream processing such as the removal of impurities and mother liquor by washing, dissolution or reaction of the crystals, and transporting or storing crystals.

Crystal size distributions may be expressed by: (1) histograms, which are the amount or fraction of mass or

TABLE II Sieved KNO₃ Crystals from Hypothetical 1-Liter Sample

Sieve no., <i>i</i>	<i>L</i> (μm)	Δ <i>M_i</i> (g/L)	<i>M(L)</i> (g/L)	<i>F(L)</i> (frac)	$\bar{L}_i$ (μm)	Δ <i>N_i</i> (no./L)	<i>N(L)</i> (frac)	<i>m</i> (g/μm·L)	<i>n</i> (no./μm·L)
1	707	0	32.974	1.000		1611			
2	500	7.296	25.678	0.779	603.5	16	1595	0.0352	0.076
3	354	11.512	14.166	0.430	427.0	70	1525	0.0789	0.480
4	240	9.011	5.154	0.156	297.0	163	1362	0.0790	1.430
5	177	3.145	2.009	0.061	208.5	164	1198	0.0499	2.610
6	125	1.322	0.687	0.021	151.0	182	1016	0.0254	3.500
7	88	0.462	0.225	0.007	106.5	181	834	0.0125	4.900
8	63	0.159	0.066	0.002	75.5	175	659	0.0064	7.000
9	44	0.055	0.011	0.000	53.5	171	488	0.0029	9.000
10	0	0.011	0.000	0.000	22.0	488	0	0.0002	11.100
Total		32.974				1611			

number over each increment in size; (2) cumulative distributions, which are the total or fraction of mass or number below (or above) a given size; and (3) density functions, which are the derivatives (with respect to size) of cumulative distributions. These definitions will be illustrated by considering a hypothetical potassium nitrate system from which a 1-liter slurry sample has been withdrawn, filtered, washed, dried, and sieved to give the results shown in Table II.

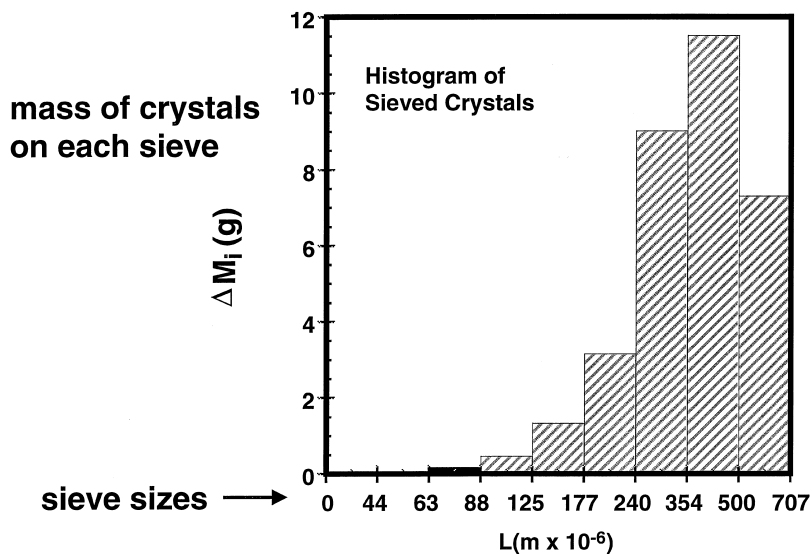
The first three and the sixth columns give the sieve data and should be read as follows: All of the sieved matter passed through the 707-μm sieve, and 7.296 g remained on the 500-μm sieve and had an arithmetic average size of 603.5 μm. Similar descriptions can be given for crystals that remained on the other sieves and pan (*L* = 0). The total crystal mass recovered was 39.974 g. A histogram of the mass distribution from these data is shown in Fig. 10.

The method by which crystals are sized gives either number or mass of crystals in a given size range. The sieve analysis in the above example gives mass distributions, so that the histogram is constructed in terms of crystal mass, and a cumulative mass distribution, *M(L)*, can be defined as the mass of crystals in the sample passing through the sieve of size *L*. In other words,

$$M(L) = \sum_{L=0}^{L(i)} \Delta M_i \quad (29)$$

Such calculations give the mass of crystals below size *L*, and the results are shown in column 4 of Table II. Column 5 gives the cumulative mass fraction distribution:

$$F(L) = \frac{M(L)}{M_{\text{total}}} \quad (30)$$

**FIGURE 10** Histogram of size distribution from example.

Transforming a mass distribution to a number distribution, or vice versa, requires a relationship between the measured and desired quantities. The mass of a single crystal, m_{crys} , is related to crystal size by the volume shape factor, k_{vol} (see Eq. (19)):

$$m_{\text{crys}} = \rho k_{\text{vol}} L^3 \quad (31)$$

Consequently, the number of crystals on a sieve in the example, ΔN_i , can be estimated by dividing the total mass on sieve i by the mass of an average crystal on that sieve. If the crystals on that sieve are assumed to have a size equal to the average of the sieve through which they have passed and the one on which they are held, $\bar{L}_i = (L_{i-1} + L_i)/2$, then:

$$\Delta N_i = \frac{\Delta M_i}{\rho k_{\text{vol}} \bar{L}_i^3} \quad (32)$$

Potassium nitrate crystals have a density of $2.11 \times 10^{-12} \text{ g}/\mu\text{m}^3$, which allows for the determination of the estimated crystal numbers on each sieve in Table II. A cumulative number distribution, $N(L)$, and a cumulative number fraction distribution, $F(L)$, can be calculated using methods similar to those for calculating $M(L)$ and $W(L)$.

Mass and population densities are estimated from the respective cumulative number and cumulative mass distributions:

$$m(\bar{L}) = \frac{\Delta M_i}{\Delta L_i} \quad (33)$$

$$n(\bar{L}) = \frac{\Delta N_i}{\Delta L_i} \quad (34)$$

So that if $\Delta L_i \rightarrow 0$,

$$m(L) = \frac{dM(L)}{dL} \Rightarrow M(L) = \int_0^L m dL \quad (35)$$

and

$$n(L) = \frac{dN(L)}{dL} \Rightarrow N(L) = \int_0^L n dL \quad (36)$$

Equations (33) and (34) are used to obtain the last two columns of Table II.

In the example, all of the results are for the given sample size of 1 liter and the quantities estimated have units reflecting that basis. This basis volume is arbitrary, but use of the calculated quantities requires care in defining this basis consistently in corresponding mass and population balances. The volume of clear liquor in the sample is an alternative, and sometimes more convenient, basis.

Moments of a distribution provide information that can be used to characterize particulate matter. The j th moment of the population density function $n(L)$ is defined as:

$$m_j = \int_0^{\infty} L^j n(L) dL \quad (37)$$

From Eq. (37), it can be demonstrated that the total number of crystals, the total length, the total area, and the total volume of crystals, all in a unit of sample volume, can be evaluated from the zeroth, first, second, and third moments of the population density function. Moments of the population density function also can be used to estimate number-weighted, length-weighted, area-weighted, and volume- or mass-weighted quantities. These averages are calculated from the general expression:

$$\bar{L}_{j+1,j} = \frac{m_{j+1}}{m_j} \quad (38)$$

where $j = 0$ for a number-weighted average, 1 for a length-weighted average, 2 for an area-weighted average, and 3 for a volume- or mass-weighted average.

Crystal size distributions may be characterized usefully (though only partially) by a single crystal size and the spread of the distribution about that size. For example, the dominant crystal size represents the size about which the mass in the distribution is clustered. It is defined as the size, L_D , at which a unimodal mass density function is a maximum, as shown in Fig. 11; in other words, the dominant crystal size L_D is found where dm/dL is zero. (The data used to construct Fig. 11 are from Table II.) As the mass density is related to the population density by:

$$m = \rho k_v n L^3 \quad (39)$$

the dominant crystal size can be evaluated from the population density by:

$$\frac{d(nL^3)}{dL} = 0 \quad \text{at} \quad L = L_D \quad (40)$$

The spread of the mass-density function about the dominant size is the coefficient of variation (c.v.) of the CSD.

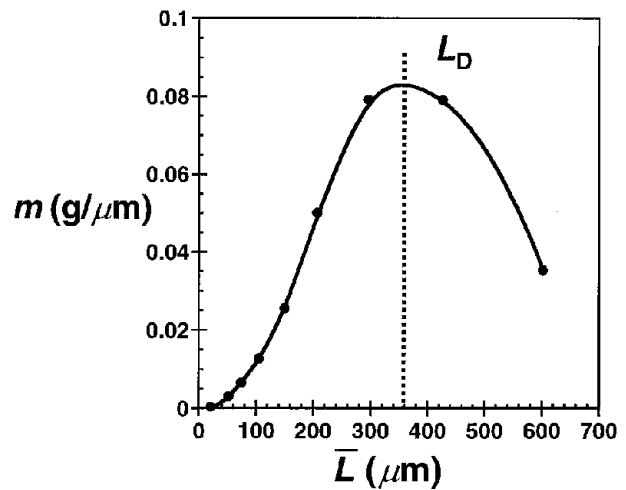


FIGURE 11 Mass-density function with single mode showing dominant size.

It is defined by:

$$\text{c.v.} = \frac{\sigma}{L_D} \quad (41)$$

and estimated from the moments of the distribution:

$$\text{c.v.} = \left[\frac{m_3 m_5}{m_4^2} - 1 \right]^{1/2} \quad (42)$$

This is especially useful for systems that cannot be described by an analytical distribution function.

V. CRYSTALLIZER CONFIGURATION AND OPERATION

Crystallization equipment can vary in sophistication from a simple stirred tank to a complicated multiphase column, and the protocol can range in complexity from simply allowing a vat of liquor to cool to the careful manipulation required of batch cyclic operations. In principle, the objectives of these systems are the same: to produce a product meeting specifications on quality at an economical yield. This section will examine some of the considerations that go into the selection of a crystallizer so as to meet these objectives.

One of the first decisions that must be made is whether the crystallizer operation is to be batch or continuous. In general, the advantages of each type of operation should be weighed in choosing one over the other, but more often the decision rests on whether the other parts of the process are batch or continuous. If they are batch, then it is likely that the crystallizer also should be batch.

The equipment required for batch crystallization can be very simple. For example, some crystalline materials are produced by simply allowing a charge of hot liquor to cool. After the crystals have formed, the magma is discharged through a filter or the liquor may be decanted and the settled slurry filtered. Very large crystals can be obtained by allowing encrustations formed on the walls of these crystallizers to grow undisturbed; after the system has come to equilibrium, the liquor is drained and the crystals are removed by scraping them from the surface.

Batch crystallizers can be used in a campaign to produce a particular product and in a second campaign to produce another product. Generally, it is not possible to operate continuous processes in this way. Batch crystallizers can handle viscous or toxic systems more easily than can continuous systems, and interruption of batch operations for periodic maintenance is less difficult than dealing with interruptions in continuous processes. The latter factor may be especially important in biological processes that require frequent sterilization of equipment. Batch crystallizers can produce a narrow crystal size distribution, whereas special processing features are required to narrow the distribu-

tion obtained from a continuous crystallizer. The effects of operating variables on crystal size distributions will be discussed in Section VI.

The throughput per unit crystallizer volume is greater for a continuous system. Batch units have several operating steps in a cycle—charging, heating or cooling, crystallizing, discharging, and cleaning—and the unit production rate is based on the total cycle time, even though the formation of crystals may occur only during a small portion of the cycle.

It may be easier to operate a continuous system so that it reproduces a particular crystal size distribution than it is to reproduce crystal characteristics from a batch unit. Moreover, the coupling of several transient variables and nucleation make it difficult to model and control the operation of a batch crystallizer.

A. Relationship of Solubility to Mode of Operation

The driving force for crystal formation can be generated through a variety of means, including cooling or heating to reduce or increase the system temperature, evaporating solvent, evaporative (flash) cooling, inducing a chemical reaction (when the reaction product is sparingly soluble, the process is called *precipitation*), adjusting pH, salting out through the addition of a nonsolvent, direct-contact cooling with a refrigerant, or some other means. All of these modes of operation can be implemented in either a batch or a continuous process. In addition, two or more of the modes may be combined to enhance the product yield.

As discussed in an earlier section, solubility is intrinsic to the solute-solvent system, and the relationship of solubility to temperature often determines the mode by which a crystallizer is operated. Recall for example that the solubility of NaCl (see Fig. 2) is essentially independent of temperature, while $\text{Na}_2\text{SO}_4 \cdot 10\text{H}_2\text{O}$ has a solubility that exhibits a strong dependence on temperature. Consequently, cooling a sodium chloride solution cannot generate significant product yield; solvent evaporation is the primary mode of NaCl production. On the other hand, reducing the temperature of a saturated solution of sodium sulfate generates substantial product and may be used alone or in combination with evaporation.

Cooling crystallizers utilize a heat sink to remove both the sensible heat from the feed stream and the heat of crystallization released or, in some cases absorbed, as crystals are formed. The heat sink may be no more than the ambient surroundings of a batch crystallizer, or (as is more likely) it may be cooling water or another process stream.

Evaporative crystallizers generate supersaturation by removing solvent from the mixture, thereby increasing the solute concentration. They may be operated under vacuum, and in those circumstances it is necessary to have a

vacuum pump or ejector as a part of the unit. If the boiling point elevation—the increase in boiling temperature due to the presence of the solute—is low, mechanical recompression of the vapor obtained from solvent evaporation may be used in some cases to produce a heat source to drive the operation.

Evaporative-cooling crystallizers are fed with a liquor whose temperature is such that solvent flashes upon feed entry to the crystallizer. They typically are operated under vacuum, and flashing of solvent increases the solute concentration in the remaining liquor while simultaneously reducing the temperature of the magma. The mode of this operation can be reduced to that of a simple cooling crystallizer by returning condensed solvent to the crystallizer body.

Salting-out crystallization operates through the addition of a nonsolvent to the magma in a crystallizer. The selection of the nonsolvent is based on the effect of the solvent on solubility, cost, properties that affect handling, interaction with product requirements, and ease of recovery. Adding a nonsolvent to the system increases the complexity of the process; it increases the volume required for a given residence time and produces a highly nonideal mixture of solvent, nonsolvent, and solute.

Melt crystallization operates with heat as a separating agent, but a crystalline product is not generated in the process. Instead, crystals formed during the operation are remelted and the melt is removed as the product. Such operations are often used to perform the final purification of products after prior separation units; for example, the purity of an acrylic acid feed may be increased from 99.5 to 99.9%. Melt crystallizers do not require solids handling units nor do they utilize solid–liquid separation equipment. Finally, in some instances the use of melt crystallization can eliminate the use of solvents, thereby reducing the environmental impact of the process.

B. Crystallizers

The basic requirements of a crystallization system are (1) a vessel to provide sufficient residence time for crystals to grow to a desired size, (2) mixing to provide a uniform environment for crystal growth, and (3) a means of generating supersaturation. Crystallization equipment is manufactured and sold by several vendors, but some chemical companies design their own crystallizers based on expertise developed within their organizations. Rather than attempt to describe the variety of special crystallizers that can be found in the marketplace, this section will provide a brief general survey of types of crystallizers that utilize the modes outlined above.

The forced-circulation crystallizer is a simple unit designed to provide high heat-transfer coefficients in either an evaporative or a cooling mode. Figure 12 shows a

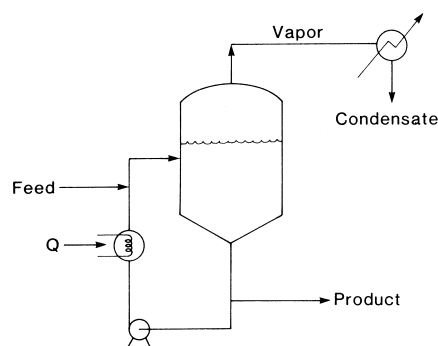


FIGURE 12 Schematic diagram of a forced-circulation evaporative crystallizer.

schematic diagram of an evaporative forced-circulation crystallizer that withdraws a slurry from the crystallizer body and pumps it through a heat exchanger. Heat transferred to the circulating magma causes evaporation of solvent as the magma is returned to the crystallizer. This type of unit is used to control circulation rates and velocities past the heat transfer surfaces, and the configuration shown is especially useful in applications requiring high rates of evaporation. A calandria that provides heat transfer through natural convection is an alternative to forced-circulation systems.

Scale formation on the heat exchanger surfaces or at the vapor–liquid surface in the crystallizer can cause operational problems with evaporative crystallizers. These can be overcome by avoiding vaporization or excessive temperatures within the heat exchanger and by properly introducing the circulating magma into the crystallizer. For example, introducing the circulating magma a sufficient distance below the surface of the magma in the crystallizer prevents vaporization upon re-entry and forces it to occur at a well-mixed zone above the point of re-entry. Alternatively, the magma may be introduced so as to induce a swirling motion that dislodges encrustations from the wall of the crystallizer at the vapor–liquid interface.

Figure 13 shows a schematic diagram illustrating the configuration of a surface cooling (indirect heat transfer) crystallizer. Heat can be transferred to a coolant in an external heat exchanger, as shown, or in coils or a jacket

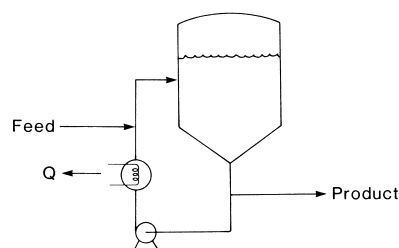


FIGURE 13 Schematic diagram of a forced-circulation, surface-cooling crystallizer.

within the crystallizer. An external cooling surface requires the use of a circulation pump, but this expense may be mitigated by obtaining a higher heat-transfer coefficient than would result with the use of coils or a jacketed vessel. The rate of heat transfer Q from the circulation loop of a cooling crystallizer must be sufficient to reduce the temperature of the feed and to remove the heat of crystallization of the solute. Assuming that no substantial crystallization occurs in the heat exchanger and limiting the difference between entering and leaving temperatures of the circulating magma ($T_{in} - T_{out}$), so as to minimize formation of encrustations, the required magma circulation rate $\dot{m}_{circ}$ can be determined from the equation:

$$\dot{m}_{circ} = \frac{Q}{[C_P(T_{in} - T_{out})]_{circ}} \quad (43)$$

where C_P is the heat capacity of the circulating magma. The methods by which Q can be evaluated were discussed in Section II. It is not uncommon to limit the decrease in magma temperature to about 3 to 5°C; therefore, both the circulation rate and heat-transfer surface must be large.

The feed to cooling crystallizers should be rapidly mixed with the magma so as to minimize the occurrence of regions of high supersaturation. Such regions lead to excessive nucleation, which is detrimental to the crystal size distribution. The type of pump used in the circulation loop also can lead to degradation of the crystal size distribution; an inappropriate pump causes crystal attrition through abrasion, fracture, or shear, and most commercial systems use specially designed axial-flow pumps that provide high flow rates and low shear.

Direct-contact refrigeration can be used if either the operating temperature of the crystallizer is low in comparison to the temperature of available cooling water or there are severe problems with encrustations. In such an operation, a refrigerant is mixed with the crystallizer contents and vaporized at the magma surface. On vaporizing, the refrigerant removes sufficient heat from the magma to cool the feed and to remove the heat of crystallization. The refrigerant vapor must be compressed, condensed, and recycled for the process to be economical. Moreover, the refrigerant must be insoluble in the liquor to minimize losses and product contamination.

Special devices for classification of crystals may be used in some applications. Figure 14 shows a draft-tube-baffle (DTB) crystallizer that is designed to provide preferential removal of both fines and classified product. As shown, feed is introduced to the fines circulation line so that any nuclei formed upon introduction of the feed can be dissolved as the stream flows through the fines-dissolution heat exchanger. The contents of the crystallizer are mixed by the impeller, which forces the slurry to flow in the indicated direction. A quiescent zone is formed between the

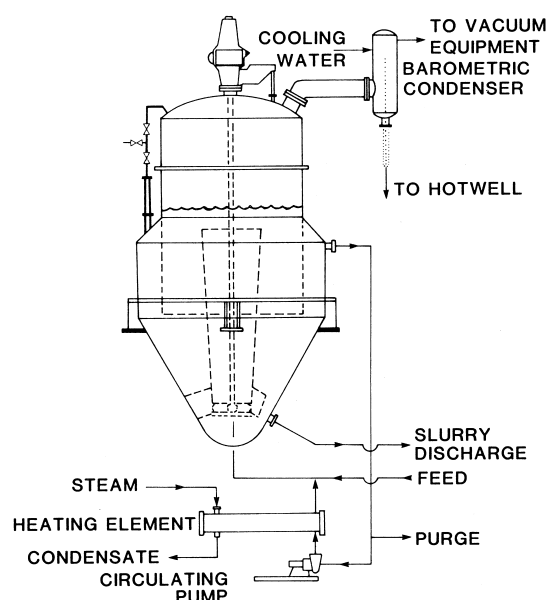


FIGURE 14 Draft-tube-baffle crystallizer. [Courtesy of Signal Swenson Division.]

baffle extending into the chamber and the outside wall of the crystallizer. Flow through the quiescent zone can be adjusted so that crystals below a certain size (determined by settling velocity) are removed in the fines-dissolution circuit. In the elutriation leg, crystals below a certain size are preferentially swept back into the crystallizer by the flow of recycled mother liquor; accordingly, larger crystals, which have a higher settling velocity, are removed preferentially from the system.

A second major type of crystallizer with special channeling devices is comprised of those having configurations like the Oslo crystallizer shown in Fig. 15. The objective of this unit is to form a supersaturated solution by evaporation in the upper chamber and to have crystal growth in the lower (growth) chamber. The use of the downflow pipe in the crystallizer provides good mixing in the growth chamber. As shown, the lower chamber has a varying diameter, which can provide some internal classification of crystals. The lowest portion of the chamber has the smallest diameter and can be considered perfectly mixed; as the chamber diameter increases, the upward velocity of the slurry decreases and larger crystals tend to settle. In principle, only small crystals are supposed to leave the chamber in the circulating slurry, to flow through the circulation pump, and to enter the upper chamber. As the probability of a crystal colliding with the impeller decreases with decreasing crystal size, the internal classification provided by the Oslo crystallizer could provide some control of contact nucleation.

Melt crystallizers can be operated in a variety of ways. In one, feed enters the crystallizer and contacts a slurry

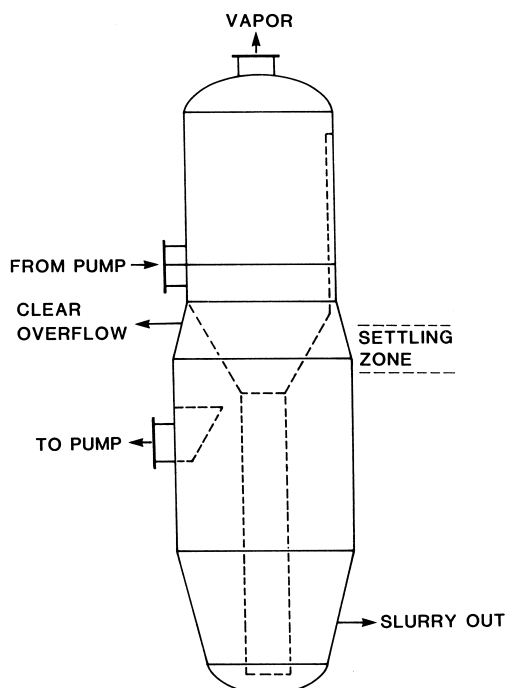


FIGURE 15 Oslo crystallizer.

of mother liquor and crystals of the desired product. The crystals are forced to move in a specific direction by gravity or rotating blades. As they flow towards the appropriate end of the crystallizer, the crystals encounter a heated region and are melted. A portion of the melt is removed as product, while the remainder flows countercurrently to the crystals, thereby providing some refining and removing impure adhering liquid.

In a second method of operation, the feed material is circulated through a bank of tubes, each of which has a diameter of up to about 8 cm. The walls of the tubes are cooled, and material crystallizes on them throughout a fixed operating period. At the end of that period, the remaining liquid is sent to a holding tank for further processing, and then the tubes are heated slowly to cause partial melting of the adhering solids. This step is known as “sweating,” and the impure “sweated” liquid produced is removed from the crystallizer and held for further processing. Finally, the product is obtained by adding additional heat to the tubes and melting the remaining adhering solids. The actual sequencing of these steps and the reprocessing of residual and sweated liquids may be quite complicated.

VI. POPULATION BALANCES AND CRYSTAL SIZE DISTRIBUTIONS

A balance on the population of crystals in a crystallizer can be used to relate process variables to the crystal size

distribution of the product or intermediate material. Such balances are not independent of those on mass and energy, and their solution requires an independent expression for nucleation kinetics.

In formulating a population balance, crystals are assumed sufficiently numerous for the population distribution to be treated as a continuous function. One of the key assumptions in the development of a simple population balance is that all crystal properties, including mass (or volume), surface area, and so forth are defined in terms of a single crystal dimension referred to as the characteristic length. For example, Eq. (19) relates the surface area and volume of a single crystal to a characteristic length L . In the simple treatment provided here, shape factors are taken to be constants. These can be determined by simple measurements or estimated if the crystal shape is simple and known—for example, for a cube $k_{\text{area}} = 6$ and $k_{\text{vol}} = 1$.

The beginning point for any balance is the following statement:

$$\begin{aligned} \text{input} + \text{generation} - \text{output} - \text{consumption} \\ = \text{accumulation} \end{aligned} \quad (44)$$

where each of the terms may be expressed as a rate or an amount. In a population balance, the number of entities (such as crystals) is the balanced quantity and each of the terms has dimensions of number of crystals per unit time for a differential balance or number of crystals for an integral balance. The principles involved in formulating a balance are outlined in the following sections, and they provide guidance in developing corresponding balances for systems whose configurations do not conform to those described here.

A. Perfectly Mixed, Continuous Crystallizers

The balance equation must be constructed for a control volume, which for a perfectly mixed crystallizer may be assumed to be the total volume of the crystallizer V_T . Then, a balance on the number of crystals in any size range (say, L_1 to $L_2 = L_1 + \Delta L$) must account for crystals that enter and leave that size range by: (1) convective flow, (2) crystal growth, (3) crystal agglomeration, and (4) crystal breakage. Agglomeration and breakage can be detected through careful inspection of product particles, and they can be quite significant in some processes. For simplicity, however, they will be assumed negligible in the present analysis. The rate of crystal growth G will be defined as in Eq. (26); i.e., the rate of change of the characteristic crystal dimension L :

$$G = \frac{dL}{dt}$$

Then,

$$\text{growth rate into the size range} = V_T(Gn)_{L_1} \quad (45)$$

$$\text{growth rate out of the size range} = V_T(Gn)_{L_2} \quad (46)$$

$$\text{removal rate of crystals in the size range} = V_{\text{out}} \int_{L_1}^{L_2} n dL \quad (47)$$

$$\text{feed rate of crystals in the size range} = V_{\text{in}} \int_{L_1}^{L_2} n_{\text{in}} dL \quad (48)$$

$$\text{accumulation rate in the crystallizer} = \frac{\partial}{\partial t} \int_{L_1}^{L_2} n V_T dL \quad (49)$$

Substituting the terms from Eqs. (46) through (49) into Eq. (44) gives:

$$\begin{aligned} V_T(Gn)_{L_1} + V_{\text{in}} \int_{L_1}^{L_2} n_{\text{in}} dL \\ = V_T(Gn)_{L_2} + V_{\text{out}} \int_{L_1}^{L_2} n dL + \frac{\partial}{\partial t} \int_{L_1}^{L_2} n V_T dL \end{aligned} \quad (50)$$

Manipulation of this equation leads to

$$\frac{\partial(nG)}{\partial L} + \frac{V_{\text{out}}n}{V_T} - \frac{V_{\text{in}}n_{\text{in}}}{V_T} = -\frac{\partial n}{\partial t} \quad (51)$$

Equation (51) may be used as a starting point for the analysis of any crystallizer that has a well-mixed active volume and for which crystal breakage and agglomeration can be ignored. As an illustration of how the equation can be simplified to fit specific system behavior, suppose the feed to the crystallizer is free of crystals and that it is operating at steady state. Then, $n_{\text{in}} = 0$ and $\partial n / \partial t = 0$. Now suppose that the crystal growth is invariant with size and time; in other words, assume the system follows the McCabe ΔL law and therefore exhibits neither size-dependent growth nor growth-rate dispersion. Then,

$$\frac{\partial(nG)}{\partial L} = G \frac{\partial n}{\partial L} \quad (52)$$

Defining a mean residence time $\tau = V_T / V_{\text{out}}$ and applying the aforementioned restrictions leads to

$$G \frac{dn}{dL} + \frac{n}{\tau} = 0 \quad (53)$$

(τ is often referred to as the drawdown time to reflect the fact that it is the time required to empty the contents from the crystallizer.) Integrating Eq. (53) with the boundary condition $n = n^\circ$ at $L = 0$:

$$n = n^\circ \exp\left(-\frac{L}{G\tau}\right) \quad (54)$$

If the crystallizer has a clear feed, growth is invariant, but if the magma volume V_T is allowed to vary, the population balance gives:

$$\frac{\partial n}{\partial t} + \frac{\partial(nG)}{\partial L} + n \frac{\partial(\ln V_T)}{\partial t} + \frac{V_{\text{out}}n}{V_T} = 0 \quad (55)$$

The system model that led to the development of the last two equations is referred to as the mixed-suspension, mixed-product removal (MSMPR) crystallizer.

Under steady-state conditions, the rate at which crystals are produced by nucleation must be equal to the difference in rates at which crystals leave and enter the crystallizer. Accordingly, for a clear feed,

$$V_T B^\circ = V_{\text{out}} \int_0^\infty n dL \Rightarrow B^\circ = \frac{1}{\tau} \int_0^\infty n dL \quad (56)$$

For crystallizers following the constraints given above,

$$B^\circ = n^\circ G \quad (57)$$

For a given set of crystallizer operating conditions, nucleation and growth rates can be determined by measuring the population density of crystals in a sample taken from either the well-mixed zone of a crystallizer or the product stream flowing from that zone. Sample analyses are correlated with Eqs. (54) and (57), and nucleation and growth rates are determined from those correlations. The sample must be representative of the crystal population in the crystallizer (or leaving the well-mixed unit), and experience with such measurements is invaluable in performing this analysis properly. Figure 16 shows a plot of

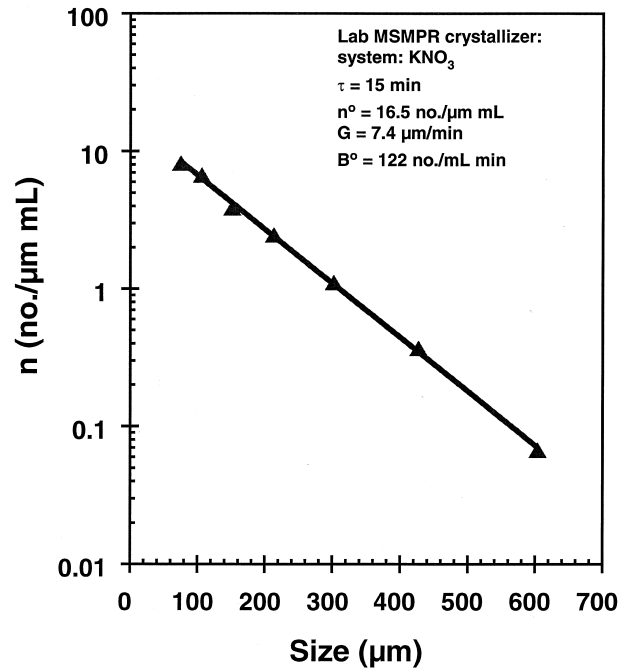


FIGURE 16 Typical population density plot from perfectly mixed, continuous crystallizer.

typical population density data obtained from a crystallizer meeting the stated assumptions. The slope of the plot of such data may be used to obtain the growth rate while the intercept can be used to estimate nucleation rate.

Many industrial crystallizers operate in a well-mixed or near well-mixed manner, and the equations derived above can be used to describe their performance. Also, the simplicity of the equations describing an MSMR crystallizer make experimental equipment configured so as to meet the assumptions leading to Eq. (54) useful in determining nucleation and growth kinetics. From a series of runs at different operating conditions, correlations of nucleation and growth kinetics with appropriate process variables can be obtained (see, for example, the discussions of Eqs. (18) and (25)). The resulting correlations can then be used to guide either crystallizer scale-up or the development of an operating strategy for an existing crystallizer.

It is often very difficult to measure supersaturation, especially in systems that have high growth rates. Even though the supersaturation in such systems is so small that it can be neglected in writing a solute mass balance, it is important in setting nucleation and growth rates. In such instances it is convenient to substitute growth rate for supersaturation by combining Eqs. (18) and (25). This gives:

$$B^\circ = k_{\text{nuc}} G^i M_T^j N^k \quad (58)$$

The constant k_{nuc} depends on process variables other than supersaturation, magma density, and intensity of mixing; these include temperature and presence of impurities. If sufficient data are available, these variables may be separated from the constant by adding more terms in a power-law correlation. k_{nuc} is specific to the operating equipment and not transferable from one equipment scale to another. The system-specific constants i and j are obtainable from experimental data and may be used in scale-up, although j may vary considerably with mixing conditions.

As shown by Eq. (54), growth rate G can be obtained from the slope of a plot of the log of population density against crystal size; nucleation rate B° can be obtained from the same data by using the relationship given by Eq. (57), with n° being the intercept of the population density plot. Nucleation rates obtained by these procedures should be checked by comparison with values obtained from a mass balance (see the later discussion of Eq. (66)).

The perfectly mixed crystallizer of the type described in the preceding discussion is highly constrained. Alteration of the characteristics of crystal size distributions produced by such systems can be accomplished only by modifications of the nucleation and growth kinetics of the system being crystallized. Indeed, examination of Eq. (54) shows that once nucleation and growth kinetics

are fixed, the crystal size distribution is determined in its entirety. In addition, such distributions have the following characteristics:

- Mass density function (from Eq. (39)):

$$m = \rho k_{\text{vol}} n^\circ L^3 \exp\left(-\frac{L}{G\tau}\right) \quad (59)$$

- Dominant crystal size (from Eq. (40)):

$$L_D = 3G\tau \quad (60)$$

- Moments of n (from Eq. (37)):

$$m_i = i! n^\circ (G\tau)^{i+1} \quad (61)$$

- Total number of crystals per unit volume:

$$N_T = \int_0^\infty n dL = m_0 = n^\circ G\tau \quad (62)$$

- Total length of crystals per unit volume:

$$L_T = \int_0^\infty nL dL = m_1 = n^\circ (G\tau)^2 \quad (63)$$

- Total surface area of crystals per unit volume:

$$A_T = k_{\text{area}} \int_0^\infty nL^2 dL = k_{\text{area}} m_2 = 2k_{\text{area}} n^\circ (G\tau)^3 \quad (64)$$

- Total solids volume per unit volume:

$$V_{\text{TS}} = k_{\text{vol}} \int_0^\infty nL^3 dL = k_{\text{vol}} m_3 = 6k_{\text{vol}} n^\circ (G\tau)^4 \quad (65)$$

- The coefficient of variation of the mass density function (from Eq. (42)) is 50%.
- The magma density M_T (mass of crystals per unit volume of slurry or liquor) is the product of the crystal density, the volumetric shape factor, and the third moment of the population density function:

$$M_T = 6\rho k_{\text{vol}} n^\circ (G\tau)^4 \quad (66)$$

System conditions often allow for the measurement of magma density, and in such cases it should be used as a constraint in evaluating nucleation and growth kinetics from measured population densities. This approach is especially useful in instances of uncertainty in the determination of population densities from sieving or other particle sizing techniques.

B. Preferential Removal of Crystals

As indicated above, crystal size distributions produced in a perfectly mixed crystallizer are constrained by the nature of the system. This is because both liquor and solids

have the same residence time distributions, and it is the crystal residence time distribution that gives the population density function the characteristic exponential form in Eq. (54). Nucleation and growth kinetics can influence the population density function, but they cannot alter the form of the functional dependence of n on L .

Crystallizers are made more flexible by the introduction of selective removal devices that alter the residence time distributions of materials flowing from the crystallizer. Three removal functions—clear-liquor advance, classified-fines removal, and classified-product removal—and their idealized removal devices will be used here to illustrate how design and operating variables can be manipulated to alter crystal size distributions. Idealized representations of the three classification devices are illustrated in Fig. 17.

Clear-liquor advance from what is called a *double draw-off crystallizer* is simply the removal of mother liquor without simultaneous removal of crystals. The primary action in classified-fines removal is preferential withdrawal from the crystallizer of crystals of a size below some specified value; this may be coupled with the dissolution of the crystals removed as fines and the return of the resulting solution to the crystallizer. Classified-product removal is carried out to remove preferentially those crystals of a size larger than some specified value. In the following discussion, the effects of each of these selective removal functions on crystal size distributions will be described in terms of the population density function n . Only the ideal solid-liquid classification devices will be examined. It is convenient in the analyses to define flow rates in terms of clear liquor. Necessarily, then, the population density function is defined on a clear-liquor basis.

Clear-liquor advance is used for two purposes: (1) to reduce the quantity of liquor that must be processed by the solid-liquid separation equipment (e.g., filter or centrifuge) that follows the crystallizer, and (2) to separate the residence time distributions of crystals and liquor. The reduction in liquor flow through the separation equipment can allow the use of smaller equipment for a fixed production rate or increased production through fixed equipment. Separating the residence time distributions of crystals and liquor means that crystals will have an average residence time longer than that of the liquor. This should, in principle, lead to the production of larger crystals, but because the crystallizer is otherwise well mixed, the crystal population density will have the same form as that for the MSMPR crystallizer (Eq. (54)).

The analysis goes as follows: Let V_{in} , V_{CL} , and V_{out} represent volumetric flow rates of clear liquor fed to the crystallizer, of clear-liquor advance, and of output slurry respectively. The population density function is given by the expression:

$$n = n^{\circ} \exp\left(-\frac{L}{G\tau_{\text{prod}}}\right) \quad (67)$$

where $\tau_{\text{prod}} = V_T / V_{out}$. Increasing V_{CL} decreases V_{out} and thereby increases the residence time of the crystals in the crystallizer. Unless the increase in magma density results in significant increases in nucleation, the utilization of clear-liquor advance will produce an increase in the dominant crystal size. Often the increase is much greater than that predicted from theory, and it is suspected that this is because the stream being removed as clear liquor actually contains varying amounts of fines. If this is the case,

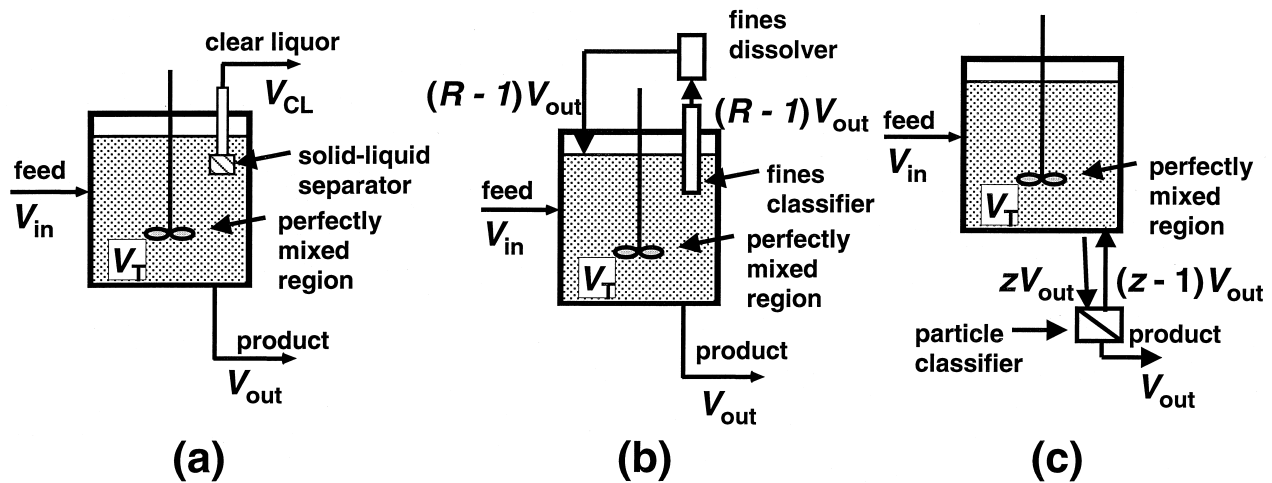


FIGURE 17 Schematic representations of idealized removal functions. (a) Clear-liquor advance, (b) classified-fines removal, (c) classified-product removal.

both clear-liquor advance and fines-removal are enhancing crystal size.

As an idealization of classified-fines removal, assume that two streams are withdrawn from the crystallizer, one corresponding to the product stream and the other a fines-removal stream. Designate the flow rate of the clear solution in the product stream to be V_{out} and set the flow rate of the clear solution in the fines-removal stream to be $(R - 1)V_{out}$. Also, assume that the device used to separate fines from the larger crystals functions so that only crystals below an arbitrary size L_F are in the fines-removal stream and that all crystals below size L_F have an equal probability of being withdrawn as fines. Under these conditions, the crystal size distribution is characterized by two mean residence times, one for the fines and the other for crystals larger than L_F , that are related by the equations:

$$\tau = \frac{V_T}{V_{out}} \quad (\text{for } L > L_F) \quad (68)$$

$$\tau_F = \frac{V_T}{RV_{out}} = \frac{\tau}{R} \quad (\text{for } L \leq L_F) \quad (69)$$

where V_T is the total volume of clear solution in the crystallizer.

For systems following invariant growth, the crystal population density in each size range will decay exponentially with the inverse of the product of growth rate and residence time. For a continuous distribution, the population densities of the classified fines and the product crystals must be the same at $L = L_F$. Accordingly, the population density for a crystallizer operating with classified-fines removal is given by:

$$n = n^\circ \exp\left[-\frac{RL}{G\tau}\right] \quad (\text{for } L \leq L_F) \quad (70)$$

$$n = n^\circ \exp\left[-\frac{(R-1)L_F}{G\tau}\right] \exp\left[-\frac{L}{G\tau}\right] \quad (\text{for } L > L_F) \quad (71)$$

Figure 18 shows how the population density function changes with the addition of classified-fines removal. The lines drawn are for a hypothetical system, but they illustrate qualitatively what can be demonstrated analytically; that is, fines removal increases the dominant crystal size, but it also increases the spread of the distribution.

A simple method for implementation of classified-fines removal is to remove slurry from a settling zone in the crystallizer. The settling zone can be created by constructing a baffle that separates the zone from the well-mixed portion of the vessel—recall, for example, the draft-tube-baffle crystallizer described in Section V—or, in small-

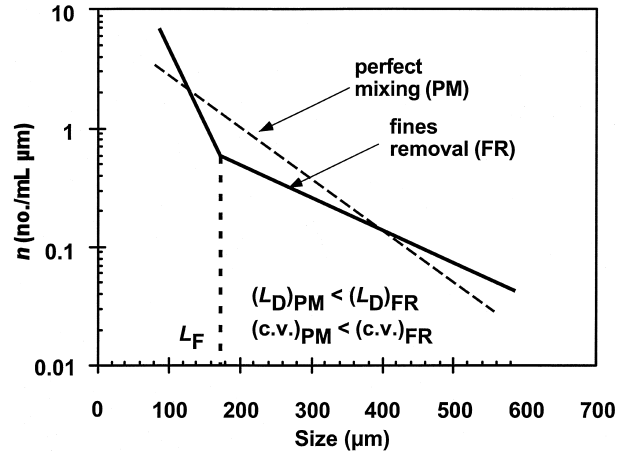


FIGURE 18 Population density plot for product from crystallizer with idealized classified-fines removal.

scale systems, by simply inserting a length of pipe or tubing of appropriate diameter into the well-mixed crystallizer chamber. The separation of crystals in the settling zone is based on the dependence of settling velocity on crystal size. Crystals entering the settling zone and having a settling velocity greater than the upward velocity of the slurry remain in the crystallizer. As the cross-sectional area of a settling zone is invariant, the flow rate of slurry through the zone determines the cut-size L_F , and it also determines the parameter R used in Eqs. (69) through (71).

In a crystallizer equipped with classified-product removal, crystals above some coarse size L_C are removed at a rate Z times the removal rate of smaller crystals. This can be accomplished by using an elutriation leg, a hydrocyclone, or a screen to separate larger crystals for removal from the system. Using the analysis of classified-fines removal as a guide, it can be shown that the crystal population density is given by the equations:

$$n = n^\circ \exp\left[-\frac{L}{G\tau}\right] \quad (\text{for } L \leq L_C) \quad (72)$$

$$n = n^\circ \exp\left[\frac{(Z-1)L_C}{G\tau}\right] \exp\left[-\frac{ZL}{G\tau}\right] \quad (\text{for } L > L_C) \quad (73)$$

where τ is defined as the residence time V_T/V_{out} . Figure 19 shows the effects of classified-product removal on crystal size distribution; the dominant crystal size is reduced and the spread of the distribution becomes narrower. Note that it is impossible for crystals smaller than L_C to leave the idealized classified-product crystallizer illustrated in Fig. 17c. Accordingly, the population densities shown on Fig. 19 for the classified-product crystallizer represent conditions *inside* the perfectly mixed region of the unit.

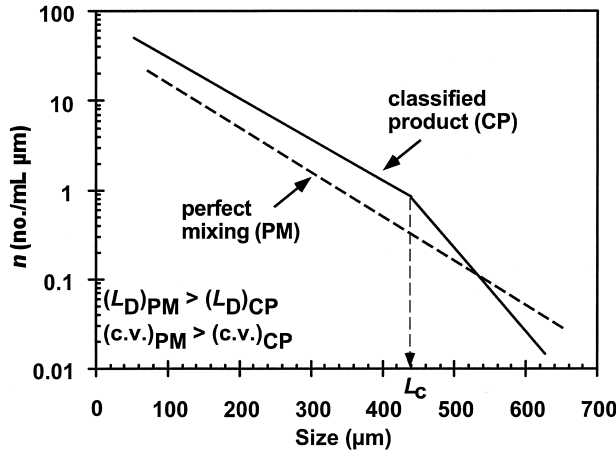


FIGURE 19 Population density plot for crystals in crystallizer with idealized classified-product removal.

If both fines and product are removed on a classified basis, the population density will be given by the equations:

$$n = n^\circ \exp\left[-\frac{RL}{G\tau}\right] \quad (\text{for } L \leq L_F) \quad (74)$$

$$n = n^\circ \exp\left[-\frac{(R-1)L_F}{G\tau}\right] \exp\left[-\frac{L}{G\tau}\right] \quad (\text{for } L_F < L < L_C) \quad (75)$$

$$n = n^\circ \exp\left[-\frac{(R-1)L_F}{G\tau}\right] \exp\left[\frac{(Z-1)L_C}{G\tau}\right] \times \exp\left[-\frac{ZL}{G\tau}\right] \quad (\text{for } L \geq L_C) \quad (76)$$

Selection of a crystallizer that has both classified-fines and classified-product removal is done to combine the best features of each: increased dominant size and narrower distribution. Figure 20 illustrates the effects of both removal functions on population density. Note that this plot of population density results from sampling the magma within a crystallizer, not from sampling the product stream, which for the ideal classification devices considered here can only have crystals larger than L_C . As discussed earlier for the classified-product crystallizer, the population densities shown in Fig. 20 represent those found in the crystallizer.

The model of the crystallizer and selective removal devices that led to Eqs. (74) through (76) is referred to as the R-Z crystallizer. It is an obvious idealization of actual crystallizers because of the perfect cuts assumed at L_F and L_C . However, it is a useful approximation to many systems and it allows qualitative analyses of complex operations.

Although many commercial crystallizers operate with some form of selective crystal removal, such devices can be difficult to operate because of fouling of heat-exchanger

surfaces or blinding of screens. In addition, classified-product removal can lead to cycling of the crystal size distribution. Often such behavior can be minimized or even eliminated by increasing the fines-removal rate.

Moments of the population density function given by Eqs. (74) through (76) can be evaluated in piecewise fashion:

$$m_i = \int_0^{L_F} L^i n dL + \int_{L_F}^{L_C} L^i n dL + \int_{L_C}^{\infty} L^i n dL \quad (77)$$

Equation (77) is used to estimate the moments of the population density function within the crystallizer, not of the product distribution. (Recall that moments of the distribution within the crystallizer are often required for kinetic equations.) Assuming perfect classification, moments of the product distribution can be obtained from the expression:

$$m_{i,\text{prod}} = \int_{L_C}^{\infty} L^i n dL \quad (78)$$

Moments can be used to characterize the material produced from or contained in a crystallizer with classified-fines or classified-product removal or to evaluate the effect of these selective removal functions on product characteristics. All that is required is the use of the equations derived earlier to relate special properties, such as coefficient of variation to the operational parameters R and Z .

C. Batch Crystallization

As with continuous crystallizers, the mode by which supersaturation is generated affects the crystal yield and size distribution; however, it is the *rate* at which such supersaturation is generated that is most important in determining product characteristics. Furthermore, there are infinite

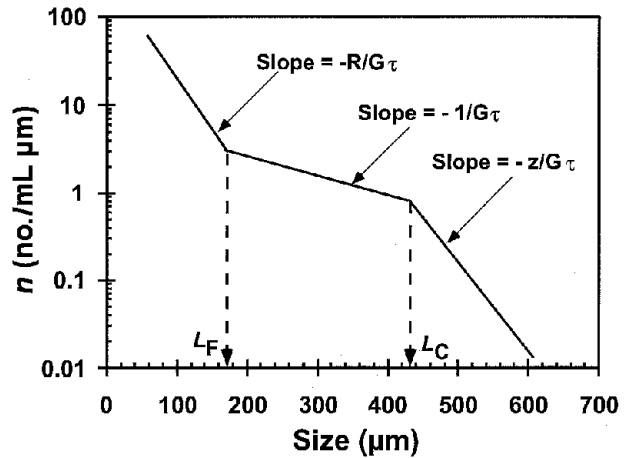


FIGURE 20 Population density plot for crystals in crystallizer with idealized classified-fines and classified-product removal.

possibilities in selecting cooling profiles, $T(t)$, or vapor generation profiles, $V(t)$, or time dependencies of precipitant or nonsolvent addition rates.

For illustrative purposes, consider that the protocol for a cooling crystallizer can involve either natural cooling—cooling resulting from exposure of the crystallizer contents to a heat sink without intervention of a control system—or manipulation of cooling to reduce the system temperature in a specific manner. In both cases, the instantaneous heat-transfer rate is given by:

$$Q = UA(T - T_{\text{sink}}) \quad (79)$$

where U is a heat-transfer coefficient, A is the area available for heat transfer, T is the temperature of the magma, and T_{sink} is the temperature of the cooling fluid. If T_{sink} is a constant, the maximum heat-transfer rate and, therefore, the highest rate at which supersaturation is generated are at the beginning of the process. This protocol can lead to excessive primary nucleation and the formation of encrustations on the heat-transfer surfaces.

The objective of programmed cooling is to control the rate at which the magma temperature is reduced so that supersaturation remains constant at some prescribed value, usually below the metastable limit associated with primary nucleation. Typically the batch is cooled slowly at the beginning of the cycle and more rapidly at the end. An analysis that supports this approach is presented later. In size-optimal cooling, the objective is to vary the cooling rate so that the supersaturation in the crystallizer is adjusted to produce an optimal crystal size distribution.

Protocols similar to those described above for cooling crystallizers exist for crystallization modes involving evaporation of solvent and the rate at which a non solvent or a reactant is added to a crystallizer.

A population balance can be used to follow the development of a crystal size distribution in batch crystallizer, but both the mathematics and physical phenomena being modeled are more complex than for continuous systems at steady state. The balance often utilizes the population density defined in terms of the total crystallizer volume, rather than on a specific basis: $\bar{n} = nV_T$. Accordingly, the general population balance given by Eq. (51) can be modified for a batch crystallizer to give:

$$\frac{\partial(nV_T)}{\partial t} + \frac{\partial(GnV_T)}{\partial L} = \frac{\partial \bar{n}}{\partial t} + \frac{\partial(G\bar{n})}{\partial L} = 0 \quad (80)$$

The solution to this equation requires both an initial condition ($\bar{n}$ at $t = 0$) and a boundary condition (usually obtained by assuming that crystals are formed at zero size):

$$\bar{n}(0, t) = \bar{n}^\circ(t) = \frac{B^\circ(t)}{G(0, t)} \quad (81)$$

The identification of an initial condition associated with the crystal size distribution is very difficult. If the system is seeded, the initial condition becomes:

$$\bar{n}(L, 0) = \bar{n}_{\text{seed}}(L) \quad (82)$$

where $\bar{n}_{\text{seed}}$ is the population density function of the seed crystals. If the system is unseeded, the nuclei often are assumed to form at size zero.

The rate of cooling, or evaporation, or addition of diluent required to maintain specified conditions in a batch crystallizer often can be determined from a population-balance model. Moments of the population density function are used in the development of equations relating the control variable to time. As defined earlier, the moments are

$$m_i = \int_0^\infty L^i \bar{n} dL \quad (83)$$

Recognizing that the zeroth moment is the total number of crystals in the system, it can be shown that:

$$\frac{dm_0}{dt} = \bar{n}^\circ G = B^\circ = \frac{dN_T}{dt} \quad (84)$$

Moment transformation of Eq. (80) leads to the following relationship:

$$\frac{\partial m_j}{\partial t} = j G m_{j-1} \quad (85)$$

Combining Eq. (85) with the relationships of moments to distribution properties developed in Section VI.A for $j = 1, 2, 3$ gives:

$$\frac{dm_1}{dt} = G m_0 \xrightarrow{m_0=N_T} \frac{dL_T}{dt} = G N_T \quad (86)$$

$$\frac{dm_2}{dt} = 2G m_1 \xrightarrow{m_1=L_T} \frac{dA_T}{dt} = 2G k_{\text{area}} L_T \quad (87)$$

$$\frac{dm_3}{dt} = 3G m_2 \xrightarrow{k_{\text{area}} m_2=A_T} \frac{dM_T}{dt} = 3G \rho \left(\frac{k_{\text{vol}}}{k_{\text{area}}} \right) A_T \quad (88)$$

where N_T is the total number of crystals, L_T is total crystal length, A_T is total surface area of the crystals, and M_T is the total mass of crystals in the crystallizer. In addition to a population balance, a solute balance must also be satisfied:

$$\frac{d(V_T C)}{dt} + \frac{dM_T}{dt} = 0 \quad (89)$$

where V_T is the total volume of the system, and C is solute concentration in the solution.

The above equations can be applied to any batch crystallization process, regardless of the mode by which supersaturation is generated. For example, suppose a model is needed to guide the operation of a seeded batch crystallizer so that solvent is evaporated at a rate that gives

a constant crystal growth rate G and no nucleation; in other words, supersaturation is to be held constant and only those crystals added at the beginning of the run are in the crystallizer. Model development proceeds as follows: combining the solute balance, Eq. (89), with Eq. (88),

$$\frac{d(V_T C)}{dt} + \frac{3\rho A_T k_{vol} G}{k_{area}} = 0 \quad (90)$$

Recognizing that the process specification requires C to be a constant and taking the derivative of Eq. (90):

$$C \frac{d^2 V_T}{dt^2} + 3\rho \left(\frac{k_{vol}}{k_{area}} \right) G \frac{dA_T}{dt} = 0 \quad (91)$$

↓ Eq. (87)

$$C \frac{d^2 V_T}{dt^2} + 6\rho k_{vol} G^2 L_T = 0 \quad (92)$$

Taking the derivative of the last equation:

$$C \frac{d^3 V_T}{dt^3} + 6\rho k_{vol} G^2 \frac{dL_T}{dt} = 0 \quad (93)$$

↓ Eq. (86)

$$C \frac{d^3 V_T}{dt^3} + 6\rho k_{vol} G^3 N_T = 0 \quad (94)$$

Suppose that the batch crystallizer is seeded with a mass of crystals with a uniform size of $\bar{L}_{seed}$. The number of seed crystals is N_{seed} , and, as the operation is to be free from nucleation, the number of crystals in the system remains the same as the number of seed crystals. The initial values of total crystal length, total crystal surface area, total crystal mass, and system volume are

$$L_T(0) = N_{seed} \bar{L}_{seed} \quad (95)$$

$$A_T(0) = k_{area} N_{seed} \bar{L}_{seed}^2 \quad (96)$$

$$M_T(0) = \rho k_{vol} N_{seed} \bar{L}_{seed}^3 \quad (97)$$

$$V_T(0) = V_{T0} \quad (98)$$

On integrating Eq. (94), the following dependence of system volume on time can be obtained:

$$C(V_{T0} - V_T) = k_{vol} \rho N_{seed} [(Gt)^3 + 3(Gt)^2 \bar{L}_{seed} + 3(Gt) \bar{L}_{seed}^2] \quad (99)$$

Therefore, for the specified conditions, the evaporation rate ($-dV_T/dt$) is a parabolic (second-order) function of time, and the rate of heat input to the crystallizer must be controlled to match the conditions called for by Eq. (99).

If a cooling mode is used to generate supersaturation, an analysis similar to that given above can be used to derive

an appropriate dependence of system temperature on time. The result depends upon the relationship of solubility to temperature. If that relationship is linear, the cooling rate varies with time in a parabolic manner; i.e.,

$$-\frac{dT}{dt} = C_1 t^2 + C_2 t + C_3 \quad (100)$$

An approximation to the temperature–time relationship that serves as a good starting point for establishing a fixed protocol is given by:

$$T = T_0 - (T_0 - T_{final}) \left(\frac{t}{\tau} \right)^3 \quad (101)$$

where τ is the overall batch run time.

It is clear that stringent control of batch crystallizers is critical to obtaining a desired crystal size distribution. It is also obvious that the development of a strategy for generating supersaturation can be aided by the types of modeling illustrated above. However, the initial conditions in the models were based on properties of seed crystals added to the crystallizer. In operations without seeding, initial conditions are determined from a model of primary nucleation.

D. Effects of Anomalous Growth

Throughout this section, crystals have been assumed to grow according to the McCabe ΔL law. This has simplified the analyses of both continuous and batch crystallizers and, indeed, crystal growth often follows the ΔL law. However, as outlined in Section III, size-dependent growth and growth-rate dispersion contribute to deviations from the models developed here. Both of these phenomena lead to similar results: In continuous, perfectly mixed crystallizers, the simple expression for population density given by Eq. (54) is no longer valid. Both size-dependent growth and growth-rate dispersion due to the existence of a random distribution of growth rates among crystals in a magma lead to curvature in plots of $\ln n$ vs. L . Models for both causes of this behavior exist but are considered beyond the scope of the present discussion. In batch crystallization, the effects of anomalous growth lead to a broadening of the distribution, as was illustrated in Fig. 6.

E. Summary

The discussion presented here has focused on the principles associated with formulating a population balance and applying simplifying conditions associated with specific crystallizer configurations. The continuous and batch systems used as examples were idealized so that the principles

could be illustrated, but the concepts can be applied to more complicated configurations. Additionally, there has been a growing body of work on aspects of population balance formulation that greatly extends the ability to describe complex systems. Such work has involved anomalous crystal growth, crystal agglomeration, and crystal breakage and necessarily results in substantially more complex models.

SEE ALSO THE FOLLOWING ARTICLES

CRYSTAL GROWTH • CRYSTALLOGRAPHY • PRECIPITATION REACTIONS • SEPARATION AND PURIFICATION OF BIOCHEMICALS • SOLID-STATE CHEMISTRY • X-RAY ANALYSIS

BIBLIOGRAPHY

- Moyers, G. C., and Rousseau, R. W. (1986). In "Handbook of Separation Process Technology" (R. W. Rousseau, ed.), Wiley, New York.
- Mullin, J. W. (1993). "Industrial Crystallization," 3rd ed. Butterworth-Heinemann, London.
- Myerson, A. S. (1993). "Handbook of Industrial Crystallization," Butterworth-Heinemann, London.
- Randolph, A. D., and Larson, M. A. (1988). "Theory of Particulate Processes," 2nd ed. Academic Press, San Diego, CA.
- Rousseau, R. W. (1993). In "Kirk-Othmer Encyclopedia of Chemical Technology," Vol. 7, 4th ed., pp. 683–730, John Wiley & Sons, New York.
- Rousseau, R. W. (1997). In "Encyclopedia of Separation Technology," Vol. 1 (D. M. Ruthven, ed.), pp. 393–439, Wiley Interscience, New York.
- Tavare, N. S. (1995). "Industrial Crystallization: Process Simulation, Analysis and Design," Plenum, New York.



Distillation

M. R. Resetarits

Koch-Glitsch, Inc.

M. J. Lockett

Praxair, Inc.

- I. Distillation Equipment
- II. Distillation Theory
- III. Distillation Column Design
- IV. Applications of Distillation
Including Energy Considerations

GLOSSARY

Azeotrope Mixture that does not change in composition on distillation and usually has a boiling point higher or lower than any of its pure constituents.

Column (tower) Vertical cylindrical vessel in which distillation is carried out.

Distillate Product of distillation formed by condensing vapor.

Efficiency (overall column efficiency) Ratio of the number of theoretical stages required to effect a distillation separation to the number of actual trays.

Height of a theoretical plate (HETP) Height of packing in a distillation column that gives a separation equivalent to one theoretical stage.

K value Ratio of the concentration of a given component in the vapor phase to its concentration in the liquid phase when the phases are in equilibrium.

Packing Specially shaped metal, plastic, or ceramic material over which the liquid trickles to give a large surface area for contact with the vapor.

Reflux ratio Ratio of the flow rate of the liquid that is returned to the top of the column (the reflux) to the flow rate of the overhead product.

Relative volatility Ratio of the K values of two components; a measure of the ease with which the two components can be separated by distillation.

Theoretical stage Contact process between vapor and liquid such that the exiting vapor and liquid streams are in equilibrium.

Trays (plates) Perforated metal sheets, spaced at regular intervals within a column, on which intimate contact of vapor and liquid occurs.

Vapor pressure Pressure at which a liquid and its vapor are in equilibrium at a given temperature.

DISTILLATION is a physical process for the separation of liquid mixtures that is based on differences in the boiling points of the constituent components. The art of distillation is believed to have originated in China around 800 BC. Early applications of the process were concerned



FIGURE 1 Large distillation column for the production of styrene. [Courtesy of Shell Chemical Canada, Ltd.]

with alcoholic beverage production and the concentration of essential oils from natural products. Over the centuries the technique spread widely, and the first book on the subject, *Das kleine Destillierbuch*, by Brunswig, appeared in 1500. Originally, distillation was carried out in its simplest form by heating a liquid mixture in a still pot and condensing the vapor that boiled off. Condensation was simply carried out by air cooling and later in water-cooled condensers. The origin of the word *distillation* is the Latin *destillare*, which means “dripping down,” and it is related to the dripping of condensed vapor product from the condenser.

I. DISTILLATION EQUIPMENT

A. General Description

Distillation is the dominant separation process in the petroleum and chemical industries. It is carried out continuously more often than batchwise, in large, vertical, hollow cylindrical columns (or towers). [Figure 1](#) shows a large distillation column with its associated piping, heat exchangers, vessels, ladders, platforms, and support structures. [Figure 2](#) shows a simple schematic representation.

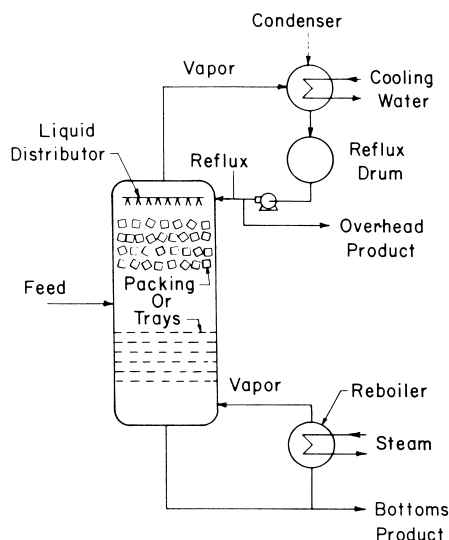


FIGURE 2 Schematic representation of a distillation column.

The process of distillation begins with a feed stream that requires treatment. It is usually necessary to separate the feed into two fractions: a low-boiling fraction (the light product) and a high-boiling fraction (the heavy product). The feed can be in a vapor or a liquid state or a mixture of both. Assuming the feed in Fig. 2 is a liquid, after entering the column it flows down through a series of trays or a stack of packing (see Section 1.B). Liquid leaving the bottom of the column is split into a bottoms product and a fraction that is made available for boiling. The bottoms product, which is rich in low-volatility components, is sometimes called the *tails* or the *bottoms*. A heat exchanger (the reboiler) is employed to boil the portion of the bottoms liquid that is not drawn off as product. The vapor produced flows up through the column (through the trays or packing) and comes into intimate contact with the downflowing liquid. After the vapor reaches and leaves the top of the column, another heat exchanger (the condenser) is encountered where heat is removed from the vapor to condense it. The condensed liquid leaving the condenser passes to a reflux drum and from there is split into two streams. One is the overhead product, which is rich in high-volatility components and is usually called the *distillate* or sometimes the *make* or the *overheads*. The other liquid stream is called the *reflux* and is returned to the top of the column. As the reflux liquid flows down the column, it comes into intimate contact with upflowing vapor. Approximately halfway down the column the reflux stream meets the liquid feed stream and both proceed down the column.

The reflux liquid returned to the top of the column has a composition identical to that of the overhead product. As the reflux, which is rich in high-volatility compo-

nents, encounters upflowing vapor, which is not as rich in these components, the difference in composition, or lack of equilibrium between the two phases, causes high-volatility components to transfer from the liquid to the vapor and low-volatility components to transfer from the vapor to the liquid. The upflowing vapor is made richer in high-volatility components and vice versa for the liquid.

Refluxing improves the separation that is achieved in most distillation columns. Any reflux rate increase, however, requires an increase in the rate of vapor production at the bottom of the column and hence an increase in energy consumption.

Contrary to the implication of Fig. 2, the condenser is usually not located at the top of the column and instead is often located some 3 to 6 m above the ground on a permanent scaffold or platform. The reflux drum is located beneath the condenser. A pump sends the reflux liquid to the top of the column and the distillate to storage or further processing.

The average distillation column at a typical refinery or petrochemical plant is probably 1 to 4 m in diameter and 15 to 50 m tall. Some columns, however, are 15 m in diameter and can extend to a height of 100 m. Columns taller than this are unfeasible to construct and erect. In addition, column height-to-diameter ratios greater than 30 are uncommon because of the support problems encountered with tall, thin columns. Most distillation columns in industrial service are bolted onto thick concrete slabs. Tall, thin columns can employ guy wires for extra support when shell thicknesses are insufficient to prevent excessive sway in the face of high winds.

Elliptical or spherical heads are employed at the top and bottom of the column. Whenever possible, industrial columns are fabricated from carbon steel, but when corrosive chemicals are encountered, columns can be made from, or lined with, more expensive materials such as stainless steel, nickel, titanium, or even ceramic materials. Operation at low temperatures also requires the use of more expensive materials. Shell thickness is generally between 6 and 75 mm. Large-diameter, high-pressure columns require thick shells to prevent shell rupture. Hoop-stress considerations alone dictate a shell thickness of 70 mm for a carbon steel column that is 3 m in diameter and operating at a pressure of 35 bars. At a height of 30 m such a vessel would weigh approximately 180,000 kg. Fortunately, most distillations are run at pressures much less than 35 bars, and thinner and less expensive columns can be employed. Column height also affects shell thickness. Height increases require shell thickness increases to combat wind forces. In addition, columns that are operated below atmospheric pressure require extra shell thickness and/or reinforcement rings to prevent column deformation or collapse. Most columns are wrapped with about

TABLE I Typical Steam Pressures Available for Distillation

Designation	Pressure (bars)	Condensation temperature (°C)
Low pressure	2.5	127
Medium pressure	15	198
High pressure	40	250

75 to 150 mm of insulation to prevent heat gain or loss, since distillation fluids are often at temperatures other than ambient.

Some distillation columns must handle two or more feed streams simultaneously. Furthermore, alternative feed nozzles are often provided to allow the actual feed-point locations to be altered. By optimizing the feed-point locations, energy consumption in the reboiler can often be minimized.

The most common energy source used in reboilers is steam. Most refineries and petrochemical plants have several steam pressure levels available. Some examples are listed in Table I. The condensation temperature of the steam used in the reboiler must be approximately 15°C greater than the boiling temperature of the bottom product. Other common heat sources used in reboilers are hot oil, hot water, and direct firing by burning oil or gas. In contrast, low-temperature columns, in ethylene plants, for example, often use propylene in a refrigeration circuit as the heating and cooling medium.

B. Column Internals

Sieve trays (Fig. 3) and valve trays (Fig. 4) are the two types of distillation trays most commonly used. In recent years these have supplanted previously widely used

bubble-cap trays except when very large flow-rate range-abilities are needed. Figure 5 shows that liquid flows across the tray deck over the outlet weir and passes down the downcomer to the next tray.

Vapor passes through holes in the tray deck where it comes into contact with the liquid to form a froth, foam, or spray. Columns operating at high pressures typically must handle large volumetric liquid flow rates per unit cross-sectional column area. Under such conditions, multiple liquid flow passes are used. Figure 6 shows two- and four-pass arrangements. Compared with a single-pass tray (Fig. 5), multipass trays have more downcomer area and a longer total outlet weir length and are capable of handling higher liquid rates. However, the number of liquid flow passes is usually minimized since multipass trays are prone to liquid and vapor maldistribution and, because they are structurally more complex, they are more expensive.

Recently there has been an increasing trend to replace the conventional trays depicted in Fig. 5 by trays having receiving pans that terminate some 15 cm above the tray deck. This provides more column cross-sectional area for vapor flow and allows increased vapor capacity. Even greater vapor capacity can be obtained from trays that utilize localized, upward co-current flow of vapor and liquid. But, as each tray then requires a vapor-liquid separation device, they are more expensive and are used only in specialized applications.

As an alternative to trays, especially at low volumetric liquid-to-vapor ratios, packing can be used to promote vapor-liquid contact. One approach is to dump specially shaped pieces of metal, glass, or ceramic material into the column, wherein they are supported on a grid. An example of dumped or random packing is shown in Fig. 7.

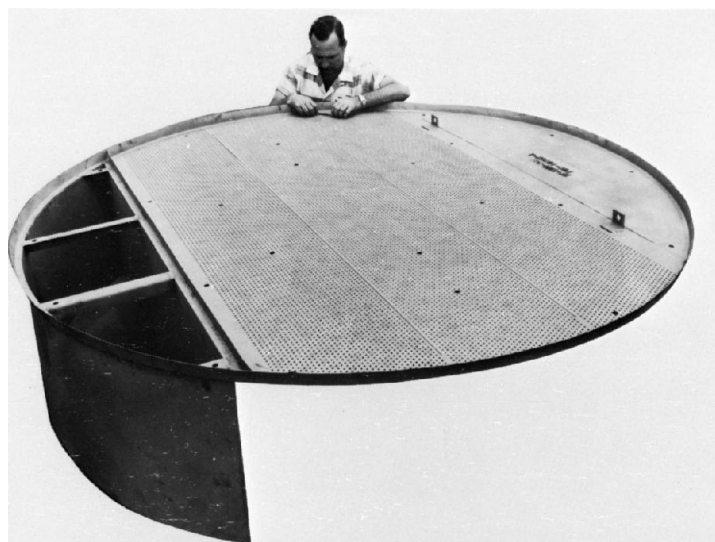


FIGURE 3 Sieve tray. [Courtesy of Koch-Glitsch, Inc.]

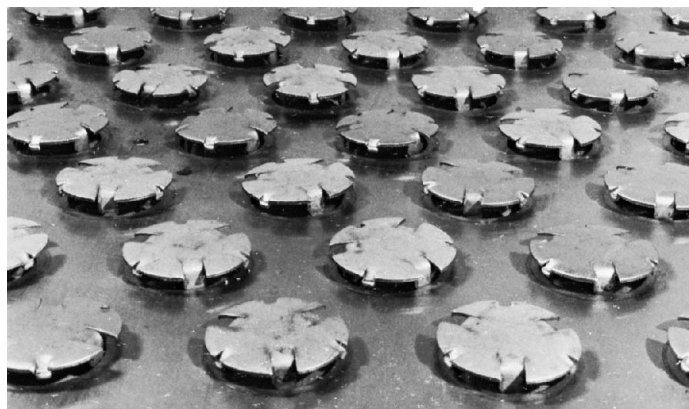


FIGURE 4 Valve tray. [Courtesy of Koch–Glitsch, Inc.]

Another approach is to fabricate and install a precisely defined packing structure, which is carefully placed to fill the column. An example of a structured packing is shown in Fig. 8. Both types of packing are most commonly made from stainless steel. The surface area per unit volume is a key variable. Large surface area packings have lower efficiencies, higher capacities, lower pressure drops, and lower costs than small surface area packings.

Liquid is introduced into a packed column via a distributor (Fig. 9), which causes a large number of liquid streams to trickle over the surface of the packing. The design of the distributor is often critical for successful packed-column operation. Structured packing generally has a higher capacity for vapor–liquid flow than dumped packing when compared under conditions of identical mass-transfer performance, but is usually more expensive. In general, packing has a lower pressure drop than trays, although it is often more expensive and less reliable in operation. Structured packing has proven to be particularly advantageous in vacuum and air separation columns.

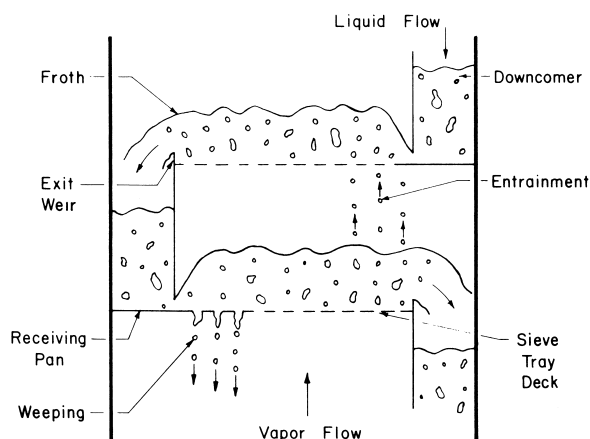


FIGURE 5 Single-pass distillation trays.

II. DISTILLATION THEORY

The process of distillation depends on the fact that the composition of the vapor that leaves a boiling liquid mixture is different from that of the liquid. Conversely, drops of liquid that condense from a vapor mixture differ in composition from the vapor.

A key physical property in distillation theory is the vapor pressure. Each pure component has a characteristic vapor pressure at a particular temperature, and vapor pressure increases with temperature and generally with a reduction in molecular weight. Vapor pressure is defined as the pressure at which a liquid and its vapor can coexist in equilibrium at a particular temperature.

The vapor pressure of a liquid mixture is given by the sum of the partial pressures of the constituents. Raoult's law is

$$p_i = p_i^\circ x_i \quad (1)$$

where p_i is the partial pressure of component i , p_i° the vapor pressure of pure component i , and x_i the mole fraction of component i in the liquid. For a vapor mixture, Dalton's law is

$$p_i = y_i \pi \quad (2)$$

where y_i is the mole fraction of component i in the vapor, and π is the total pressure. Combining Raoult's and Dalton's laws,

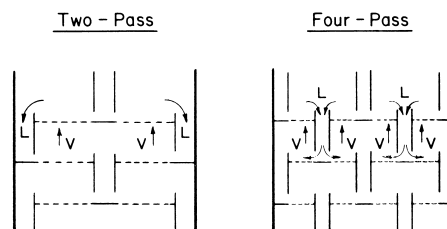


FIGURE 6 Multipass distillation trays.



FIGURE 7 Dumped packing. [Courtesy of Sulzer Chemtech, Ltd.]

$$y_i = (p_i^\circ / \pi) x_i \quad (3)$$

Equation (3) relates the composition of a liquid to the composition of its equilibrium vapor at any pressure and temperature (since p_i° depends on temperature). Equation (3) is often written:

$$y_i = K_i x_i \quad (4)$$

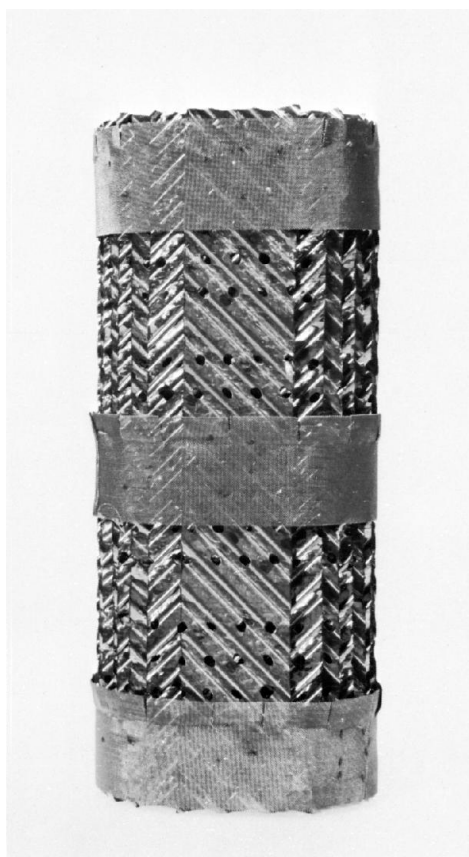


FIGURE 8 Structured packing. [Courtesy of Koch-Glitsch, Inc.]

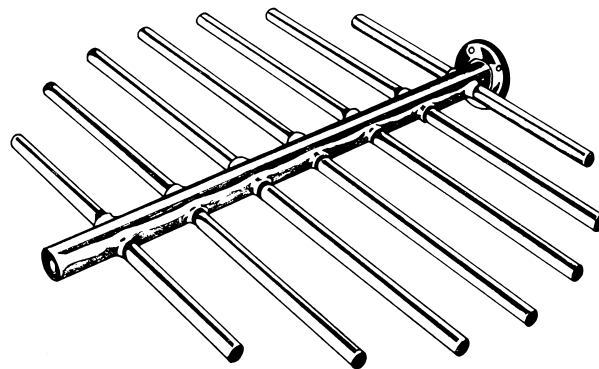


FIGURE 9 Packed column distributor.

where the equilibrium K value,

$$K_i = p_i^\circ / \pi \quad (5)$$

Mixtures that obey Eq. (5) exactly are termed *ideal mixtures*.

Deviations from ideality often occur, and the K_i value depends not only on temperature and pressure but also on the composition of the other components of the mixture. A more detailed discussion of vapor-liquid equilibrium relationships for nonideal mixtures is outside the scope of this article.

The relative volatility α of components 1 and 2 is obtained from Eq. (4) as:

$$\alpha_{12} = K_1 / K_2 = p_1^\circ / p_2^\circ = (y_1 / x_1) (x_2 / y_2) \quad (6)$$

For a binary mixture,

$$x_1 + x_2 = 1 \quad \text{and} \quad y_1 + y_2 = 1 \quad (7)$$

Substituting into Eq. (6) gives:

$$y_1 = \alpha_{12} x_1 / [1 + (\alpha_{12} - 1) x_1] \quad (8)$$

Figure 10 shows the relationship between y_1 and x_1 for different values of α_{12} calculated from Eq. (8). When two components have close boiling points, by implication they have similar vapor pressures, so that α_{12} is close to unity. Separation of mixtures by distillation becomes more difficult as α_{12} approaches unity. Figure 11 indicates some of the x, y diagrams that can be obtained for distillation systems. Also shown are corresponding temperature-composition diagrams. The saturated vapor or dewpoint curve is determined by finding the temperature at which liquid starts to condense from a vapor mixture. Similarly, the saturated liquid or bubble-point curve corresponds to the temperature at which a liquid mixture starts to boil. For ideal mixtures, the dewpoint and bubble-point curves can be calculated as follows. From Eq. (3), at the dew point, since

$$\sum_{i=1}^n x_i = 1$$

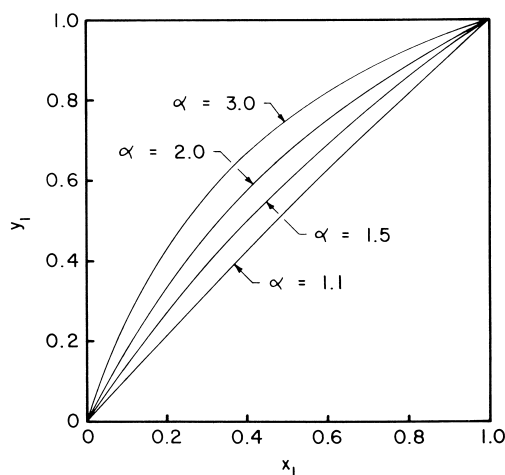


FIGURE 10 Vapor (y_1) versus liquid (x_1) concentration as a function of relative volatility.

where there are n components in the mixture,

$$\sum_{i=1}^n (y_i \pi / p_i^\circ) = 1 \quad (9)$$

Similarly, at the bubble point,

$$\sum_{i=1}^n y_i = 1$$

Therefore,

$$\sum_{i=1}^n (p_i^\circ x_i / \pi) = 1 \quad (10)$$

Since p_i° is a function of temperature, the dewpoint and bubble-point temperatures for an ideal vapor or liquid mixture can be determined as a function of the total pressure π from Eq. (9) or (10), respectively. An analogous procedure can be used for real mixtures, but the nonidealities of the liquid and vapor phases must be accounted for.

Azeotropes occur when $x_1 = y_1$, as indicated in Figs. 11c and d. Distillation of a mixture having the composition of an azeotrope is not possible since there is no difference in composition between vapor and liquid. Figure 11c shows how the azeotrope composition is affected as the pressure is changed.

When complex multicomponent mixtures are distilled, particularly those associated with oil refining, it is difficult to characterize them in terms of their components. Instead, they are characterized in terms of their boiling range, which gives some indication of the quantities of the components present. The true boiling point distillation (TBP) is probably the most useful, in which the percent distilled is recorded as a function of the boiling temperature of the mixture. For the TBP distillation, a 5 : 1 reflux ratio is often used with 15 theoretical stages in a laboratory characterization column (see Section III).

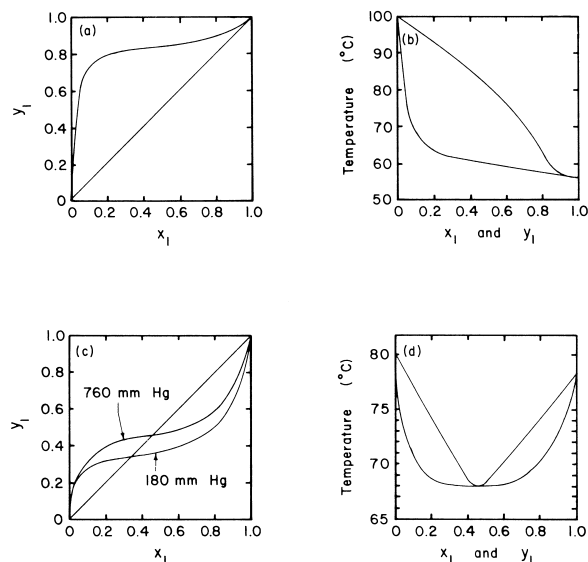


FIGURE 11 x, y and corresponding temperature-composition diagrams. (a, b) Acetone (1)–water at 1.0 bar; (c) ethanol (1)–benzene at 1.0 and 0.24 bar; (d) ethanol (1)–benzene at 1.0 bar.

III. DISTILLATION COLUMN DESIGN

It is convenient to perform calculations for both packed and trayed distillation columns in terms of theoretical equilibrium stages. A theoretical equilibrium stage is a contact process between liquid and vapor in which equilibrium is achieved between the streams leaving the theoretical stage. Figure 12 shows a representation of a theoretical stage. The compositions of y_{out} and x_{out} are in equilibrium, and the temperature and pressure of V_{out} and L_{out} are identical. The composition of y_{out} is related to x_{out} by an equilibrium relationship such as Eq. (4) or, for a binary mixture, Eq. (8). For calculation purposes, a distillation column can be modeled as a series of theoretical stages stacked one above the other. The design of a new distillation column to achieve a target separation can be broken down into a sequence of steps:

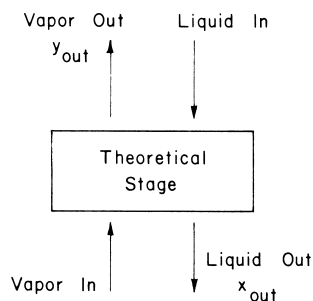


FIGURE 12 Theoretical stage concept.

1. Fix the pressure of operation of the column.
2. Determine the number of theoretical stages necessary to achieve the required separation as a function of the reflux ratio R .
3. Estimate the optimum value for R .
4. Relate the required number of theoretical stages to the actual height of the column needed.
5. Determine the necessary column diameter.
6. Refine steps 1 to 5 to achieve an optimum design.

The following sections deal with steps 1 to 5 in more detail.

A. Column Operating Pressure

The condensing temperature of the overhead vapor is reduced by lowering the column pressure. Very often, cooling water is used for condensation, and typically it has a temperature of $\sim 35^\circ\text{C}$. Consequently, the condensing vapor must have a temperature of not less than $\sim 50^\circ\text{C}$, and this sets the lower limit of the column operating pressure.

The boiling temperature of the bottoms product increases as the column pressure increases. Typically, medium-pressure steam, which has a temperature of $\sim 200^\circ\text{C}$, is used in the reboiler.

When this steam is used for heating, the bottoms product cannot have a boiling temperature greater than $\sim 185^\circ\text{C}$ which sets an upper limit on the column operating pressure.

Other heating and cooling arrangements can be employed, such as the use of a refrigerant in the condenser or higher pressure steam in the reboiler, but they increase costs and are avoided whenever possible. An additional consideration that often limits the maximum temperature of the bottoms product is polymerization and product degradation at high temperatures (and therefore at high pressures). Furthermore, at lower pressures the relative volatility tends to increase so fewer theoretical stages are required, but at the same time the column diameter tends to increase.

As a result of these factors the distillation pressure varies widely. Typically, the distillation pressure falls as the molecular weight of the feed increases. Some typical operating pressures and temperatures are shown in Table II.

B. Calculation of the Required Number of Theoretical Stages

Figure 13 shows a McCabe–Thiele diagram, which can be used when the mixture to be distilled consists of only two components or can be represented by two components. Starting at the required overhead product composition x_D , an upper-section operating line is drawn hav-

TABLE II Typical Operating Conditions in Distillation

	Pressure (bars), top	Temperature ($^\circ\text{C}$)		Theoretical stages
		Top	Base	
Demethanizer	33	−94	−8	32
Deethanizer	28	−18	72	40
Ethane–ethylene splitter	21	−29	−45	80
Propane–propylene splitter	18	45	60	150
Isobutane– <i>n</i> -butane splitter	7	45	65	60
Deisohexanizer	1.6	55	120	60
Oxygen–nitrogen separator	1.1	−194	−178	70
Ethylbenzene–styrene separator	0.06	55	115	85
Crude oil distillation	0.03	93	410	—

ing a slope $R/(R + 1)$. The operating line for the lower section below the feed is drawn by joining the required bottom composition to a point located by the intersection of the upper section operating line and the q line. The q line of Fig. 13 represents a liquid feed at its bubble point, but the slope of the q line differs for other thermal conditions of the feed. The number of theoretical stages required is determined by stepping off between the operating lines and the equilibrium line, as shown in Fig. 13. Each step on the diagram represents a theoretical stage. For the example shown, only nine theoretical stages are required, but usually many more are needed in industrial columns.

In practice, feeds rarely consist of only two components, and the McCabe–Thiele diagram cannot be used.

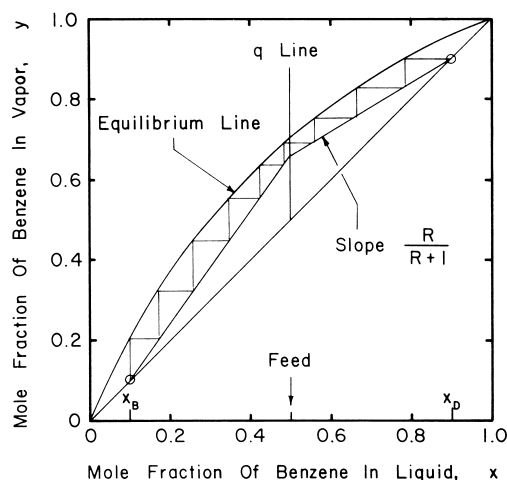


FIGURE 13 McCabe–Thiele diagram for benzene and toluene (top column pressure, 1.0 bar).

For multicomponent mixtures, the approach is to solve a complex system of matrix equations involving vapor and liquid compositions, flow rates from each theoretical stage, and temperature and pressure distributions through the column. This procedure, known as *tray counting* or *column simulation*, usually gives the required reflux ratio for specified product compositions and number of theoretical stages. Several commercial computer programs are available for tray counting.

C. Optimum Reflux Ratio

By using the procedures outlined in Section III.B, it is possible to determine the number of theoretical stages required to achieve the desired separation as a function of the reflux ratio (Fig. 14). Two limits are apparent: the minimum reflux ratio at which an infinite number of theoretical stages is necessary and the minimum number of theoretical stages that would be needed as the reflux ratio tends toward infinity. (A column operating with no feed and no product withdrawals operates at total reflux.) The optimum reflux ratio depends mainly on a balance between the investment cost of extra stages, hence extra column height, which results as R is reduced, and the operating cost of the heating medium used in the reboiler, which increases as R is increased. Generally, the optimum reflux ratio is about 1.2 to 1.5 times the minimum value.

D. Column Height

The number of actual trays required in a column can be determined from the calculated number of theoretical stages by invoking an efficiency. Various definitions of efficiency

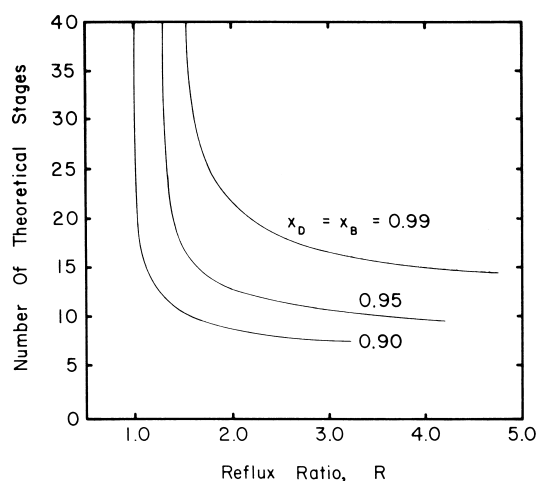


FIGURE 14 Theoretical stages versus reflux ratio (benzene–toluene at 1.0 bar). x_D , mole fraction benzene in overhead; x_B , mole fraction toluene in bottoms.

are used, but the simplest is an overall column efficiency E_o for which

$$\text{Actual trays} = \text{Theoretical stages}/E_o \quad (11)$$

For distillation, E_o is typically in the range 0.5 to 0.9. The vertical spacing between trays ranges from 200 to 900 mm. In some trayed columns, an undesirable bubbly foam can form above the liquid–vapor mixture. Antifoam chemicals must be added to such columns or diameters or tray spacings must be increased. Packed columns foam less often than trayed columns.

The required height of a packed column is determined from:

$$\text{Packed height} = \text{Theoretical stages} \times \text{HETP}$$

where HETP is the height equivalent of a theoretical plate. Note that the terms *plate*, *stage*, and *tray* tend to be used interchangeably. HETP varies with the packing size and is typically in the range of 250 to 800 mm.

E. Column Diameter

The column diameter is sized to suit the maximum anticipated rates of vapor and liquid flow through the column. Usually, the diameter is determined primarily by the vapor flow rate, and a rough estimate can be obtained from:

$$D = 4.5 Q_V^{0.5} [\rho_V / (\rho_L - \rho_V)]^{0.25} \quad (12)$$

where D is the column diameter in meters, Q_V is the vapor flow rate in cubic meters per second, and ρ_V and ρ_L are the vapor and liquid densities, respectively, in kilograms per cubic meter.

Columns operated at vapor and liquid flow rates greater than those for which they were designed become “flooded.” Unexpected foaming can also cause flooding. In a flooded column, liquid cannot properly descend against the upflowing vapor. Poor separation performance results, the overhead condensation circuit fills with process liquid, the reboiler is starved of process liquid, and the column quickly becomes inoperable.

IV. APPLICATIONS OF DISTILLATION INCLUDING ENERGY CONSIDERATIONS

A. Flash Distillation

In contrast to the description of distillation given earlier, which dealt with multistage distillation, flash distillation (Fig. 15) is carried out in a single stage. Liquid flows continuously through a heater, across a valve, and into a flash vessel. By heating the liquid and reducing its pressure across the valve, partial vaporization occurs in the flash

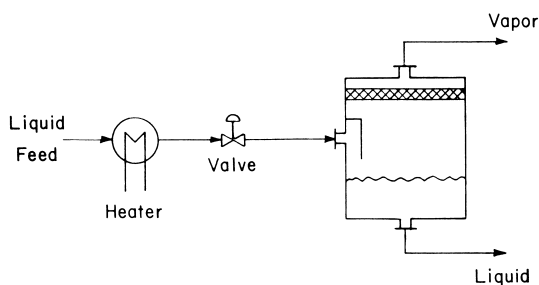


FIGURE 15 Flash distillation.

vessel. The temperature and pressure of the liquid entering the flash vessel are adjusted to achieve the required degree of vaporization. The compositions of the product streams leaving the flash vessel are different and are a function of the extent to which vaporization occurs.

Although the flash vessel itself is simple, care must be taken to ensure that the resultant vapor and liquid phases are separated completely from one another. To this end, the entering feed is often introduced tangentially rather than at a 90-degree angle to the vessel wall. An annular baffle directs the liquid droplets that are created by the flash toward the bottom of the vessel. By installing a wire mesh (approximately 75 mm thick) near the top of the vessel, fine liquid drops are prevented from leaving the top of the vessel as entrainment in the high-velocity vapor stream.

At best, only one theoretical stage is achieved by a flash distillation; however, it is used frequently in cryogenic and petroleum processing applications, where its simplicity is often attractive for nondemanding separations. Flashing often occurs in conventional distillation columns as feed and reflux streams enter. This flashing must be considered when column entrance devices and distributors are being designed.

B. Batch Distillation

Batch distillation (Fig. 16) is often preferable to continuous distillation when small quantities of feed material are processed. A liquid feed is charged to a still pot and heated until vaporization occurs. Vapor leaves the top of the column, and after condensation, part is removed as product and the rest returned to the column as reflux. As distillation proceeds, the contents of the still pot and the overhead product become richer in less volatile components. When operated at a fixed reflux ratio, an overhead product *cut* is collected until the product composition becomes unacceptable. As an alternative, the reflux ratio can be gradually increased to hold the product composition constant as the cut is taken. For a fixed rate of heat addition to the still pot, the latter option results in a steadily declining product flow rate. After the first cut, subsequent

cuts can be taken to obtain lower volatility products. Intermediate cuts of mixed composition are sometimes taken between each product cut, and these are saved and later returned to the still pot for inclusion in the next batch.

C. Extractive and Azeotropic Distillation

Conventional distillation tends to be difficult and uneconomical because of the large number of stages required when the relative volatility between the components to be separated is very low. In the extreme case, in which an unwanted azeotrope is formed, distillation past the azeotrope becomes impossible. Extractive or azeotropic distillation can sometimes be used to overcome these difficulties.

Both processes involve the addition of a new material, the *solvent*, to the mixture. The solvent is chosen so as to increase the relative volatility of the components to be separated. During extractive distillation, the solvent is generally added near the top of the column, and because it has a low volatility it is withdrawn with the product at the bottom. In azeotropic distillation, the solvent is withdrawn as an azeotrope with one or more of the components to be separated—usually in the overhead product. If the ratio of the components to be separated is different in the withdrawn azeotrope from their ratio in the feed to the column, then at least a partial separation has been achieved. In both processes it is necessary to separate the solvent from the product. This can be accomplished, for example, by distillation, solvent extraction, or even gravity settling, depending on the characteristics of the components involved.

D. Reactive Distillation

Many distillation columns reside upstream or downstream of catalytic reactors. Over the last decade, catalysts have

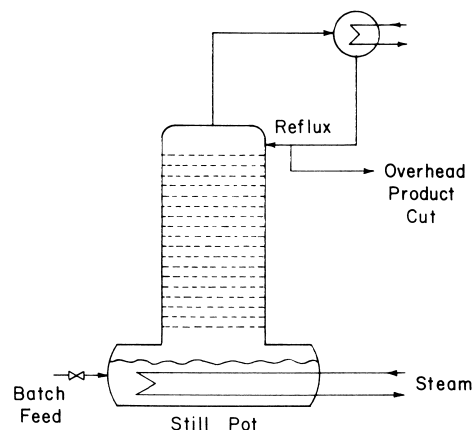


FIGURE 16 Batch distillation.

been increasingly employed inside distillation columns to simultaneously effect distillation and reaction. Oxygenates such as methyl-tert-butyl-ether (MTBE) and tertiary-methyl-ether (TAME) are produced in this manner for utilization within reformulated gasolines (RFGs). In reactive distillation, catalysts can be employed between the sheets of structured packings, on the decks or inside the downcomers of trays, or in dedicated beds between packed or trayed column sections. It is expected that reactive distillation will be used even more extensively in the future.

E. Energy Consumption

Approximately 30% of the energy used in U.S. chemical plants and petroleum refineries is for distillation, and it accounts for nearly 3% of the total U.S. annual energy consumption. The energy usage associated with some specific distillation products is shown in Table III. The cost of energy for distillation can be reduced by using waste heat such as is available from quench water in ethylene plants, for example, or exhaust steam from mechanical drivers such as compressors.

Energy costs can also be reduced by thermally linking neighboring distillation columns, as shown in Fig. 17. The overhead vapor from column 1 is condensed in an integrated condenser–reboiler, and the latent heat of condensation is used to boil the bottoms of column 2. In some cases, it may be necessary to operate columns 1 and 2 at different pressures so as to achieve the necessary temperature difference in the condenser–reboiler. The same strategy can be adopted for two columns performing identical separations in parallel. By raising the pressure of column 1, overhead vapors from column 1 can be used to drive column 2. The total energy consumption can be reduced by as much as half in this way.

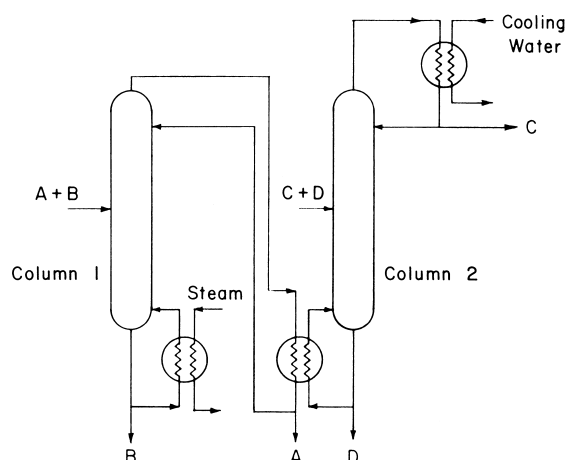


FIGURE 17 Heat-integrated columns.

TABLE III Distillation Energy Consumption

Component classification	Total U.S. distillation energy consumption (quads/yr) ^a	Specific distillation energy consumption (Btu/lb product)
Petroleum fuel fractions		
Crude distillation	0.36115	193
Vacuum distillation	0.08990	132
Catalytic hydrotreating/hydrorefining	0.07726	101
Catalytic cracking fractionator	0.06803	112
Naphtha fractionator	0.06105	132
Catalytic hydrocracking	0.05964	632
Catalytic reforming	0.04988	132
Thermal operations	0.00936	60
Ethylene primary fractionator (naphtha/gas oil cracking)	0.00205	352
Total	0.77832	331
Light hydrocarbons		
Natural gas processing	0.07495	827
Ethylene and propylene	0.04821	1517
Alkylation HF	0.04701	1046
Alkylation H ₂ SO ₄	0.03065	570
Light ends processing	0.01729	699
Isomerization	0.01312	803
Butadiene	0.01024	3151
Cyclohexane	0.00021	98
Total	0.24168	928
Water-oxygenated hydrocarbons		
Ethylene glycols	0.01065	2795
Ethanol	0.01063	9008
Phenol	0.00947	4344
Adipic acid	0.00739	4862
Methanol	0.00733	1175
Vinyl acetate (monomer)	0.00710	4797
Acetic acid	0.00701	2885
Isopropanol	0.00651	3785
Ethylene oxide	0.00554	1325
Methyl ethyl ketone	0.00481	9431
Terephthalic acid	0.00425	1756
Acetone	0.00417	2172
Dimethyl terephthalate	0.00412	1567
Formaldehyde	0.00412	733
Acetic anhydride	0.00267	1669
Propylene oxide	0.00219	1217
Glycerine	0.00202	14,870
Acetaldehyde	0.00174	1081
Total	0.10172	2366
Aromatics		
BTX ^b	0.02437	933
Styrene	0.01554	2467

continues

TABLE III (continued)

Component classification	Total U.S. distillation energy consumption (quads/yr) ^a	Specific distillation energy consumption (Btu/lb product)
Ethylbenzene	0.01388	2264
<i>o</i> -Xylene	0.00638	6019
Cumene	0.00390	1450
Total	0.06407	1515
Water-inorganics		
Sour water strippers	0.02742	240
Sodium carbonate	0.01398	1875
Urea	0.01030	133
Total	0.05170	411
Others		
Vinyl chloride (monomer)	0.01256	2188
Oxygen and nitrogen	0.00846	158
Acrylonitrile	0.00826	5434
Hexamethylenediamine	0.00612	8164
Total	0.03540	567
Remaining 30% of chemicals		
Production	0.10869	1973
Total for all component classifications	1.38158	623

From Mix, T. J., Dweck, J. S., and Weinberg, M. (1978). *Chem. Engr. Prog.* **74** (4), 49–55. Reproduced by permission of the American Institute of Chemical Engineers.

^a 1 quad = 10¹⁵ Btu.

^b Benzene-toluene-xylene.

A technique for energy reduction that has received considerable attention since 1970 is vapor recompression, or heat pumping. Vapor recompression takes advantage of the fact that when a vapor is compressed its temperature is simultaneously increased. Figure 18 shows typical temperatures and pressures associated with the use of heat pumping for splitting C₄ hydrocarbons. Through the use of a compressor, vapor leaving the top of the column is compressed from 3.8 bars and 27°C to 10.7 bars and 69°C. The compressed vapor is then hot enough to be used to boil the liquid at the bottom of the column, where the temperature is 46°C.

Vapor recompression eliminates the need for a conventional heat source, such as steam, to drive the reboiler. There is, however, an electrical energy requirement to drive the compressor which is not present in conventional distillation. The key advantage of vapor recompression is that the cost of running the compressor is often lower than the cost of driving a conventional reboiler. Under ideal conditions, the operating cost of a vapor recompression

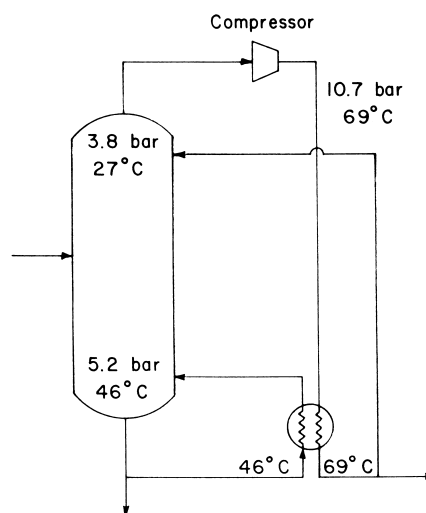


FIGURE 18 Vapor recompression.

system can be one-sixth of that associated with conventional distillation. As the temperature difference between the top and bottom of the column increases, compression costs become prohibitive. Vapor recompression is rarely used if the temperature difference exceeds 30°C.

F. Distillation Column Control

A typical control scheme for a distillation column is shown in Fig. 19. Flow controllers (FCs) regulate the flow rates of the feed and overhead products. Each flow rate is measured by a device such as an orifice plate placed upstream

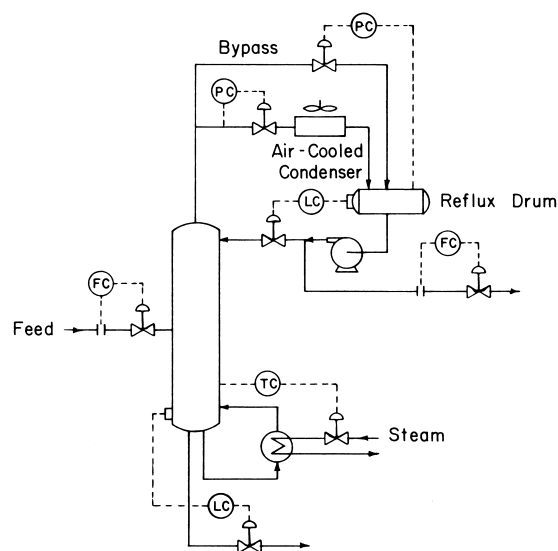


FIGURE 19 Typical distillation column control scheme.

of the control valve. The flow controller is used to open or close the control valve in response to differences between the measured flow rate and the target flow rate (the flow controller's set point). The rate of steam flow to the reboiler is regulated by measuring the temperature (usually with a thermocouple) at a point in the column and comparing this temperature to the set point of the temperature controller (TC). The rate of flow of the bottoms product is regulated by measuring the level of liquid in the column sump and opening or closing a control valve using a level controller (LC) to keep the level steady and at its set point. Similarly, the liquid level in the reflux drum is controlled by regulating the flow of reflux back to the column. Column pressure is controlled via a pressure controller (PC) acting on the condenser inlet valve, and the reflux drum pressure is controlled by a valve in the bypass line around the condenser.

The control scheme described is just one of a wide variety. In the past few years, the art and science of column control have developed rapidly, and now control system design tends to be the prerogative of the specialist control engineer.

SEE ALSO THE FOLLOWING ARTICLES

CHEMICAL THERMODYNAMICS • FLUID DYNAMICS, CHEMICAL ENGINEERING • FLUID MIXING • MEMBRANES, SYNTHETIC • PETROLEUM REFINING

BIBLIOGRAPHY

- Billet, R. (1995). "Packed Towers," VCH, Weinheim.
- Kister, H. Z. (1990). "Distillation Operation," McGraw-Hill, New York.
- Kister, H. Z. (1992). "Distillation Design," McGraw-Hill, New York.
- Lockett, M. J. (1986). "Distillation Tray Fundamentals," Cambridge University Press, Cambridge, U.K.
- Luyben, W. L. (1992). "Practical Distillation Control," Van Nostrand-Reinhold, New York.
- Seader, J. D., and Henley, E. J. (1998). "Separation Process Principles," Wiley, New York.
- Shinskey, F. G. (1984). "Distillation Control for Productivity and Energy Conservation," McGraw-Hill, New York.
- Stichlmair, J. G., and Fair, J. R. (1998). "Distillation," Wiley, New York.
- Strigle, R. F. (1994). "Packed Tower Design and Applications," Gulf Pub., Houston, TX.
- Taylor, R., and Krishna, R. (1993). "Multicomponent Mass Transfer," Wiley, New York.



Electrochemical Engineering

Geoffrey Prentice

National Science Foundation

- I. Historical Development
- II. Basic Principles
- III. Mass Transport
- IV. Current Distribution
- V. System Design

GLOSSARY

Current distribution Distribution of reaction rates on an electrode surface. Primary current distribution is calculated by considering only electric field effects; both overpotential and concentration gradients are neglected. Secondary current distribution takes both field effects and surface overpotential into account. Tertiary current distribution takes field effects, surface overpotential, and concentration gradients into account.

Current efficiency Fraction of total current that generates desired products.

Electrolytic cell Electrochemical cell that must be driven by an external power source to produce products.

Exchange current density Current density in forward and backward direction when an electrode is at equilibrium and no net current flows.

Galvanic cell Electrochemical device that converts or produces energy.

Limiting current density Maximum (diffusion-limited) current density at which a given electrode reaction can proceed. Above this limit another electrode reaction commences.

Mass-transfer boundary layer (Nernst diffusion layer) Layer adjacent to an electrode where concentrations of

reactants or products vary. Usually the thickness is of the order of 0.1–0.01 mm.

Ohmic drop Voltage loss caused by resistance of ion flow in electrolyte.

Overpotential Departure from equilibrium (reversible) potential due to passage of a net current. Concentration overpotential results from concentration gradients adjacent to an electrode surface. Surface overpotential results from irreversibilities of electrode kinetics.

Supporting (inert or indifferent) electrolyte Compounds that increase the ionic conductivity of the electrolyte but do not participate in the electrode reaction.

Wagner number Dimensionless ratio of polarization resistance to electrolyte resistance. A low value is characteristic of a primary current distribution; a high value corresponds to a secondary current distribution.

ELECTROCHEMICAL PROCESSES are employed in chemical production, metal finishing, and energy conversion. Electrochemical engineering encompasses the conception, design, scale-up, and optimization of such processes. The largest-scale electrolytic processes are aluminum and chlorine production; together they consume

over 6% of the U.S. output of electrical energy. Other commercially important processes include plating, anodizing, and electroorganic synthesis. Energy storage and conversion devices based on electrochemical principles are in widespread use. Development of electrochemical systems to reduce corrosion rates also involves electrochemical engineering. Before 1940, electrochemical engineering was practiced on an empirical basis; subsequently, it has emerged as a fundamental discipline based on the principles of thermodynamics, kinetics, fluid flow, and heat and mass transport.

I. HISTORICAL DEVELOPMENT

The discovery of electrochemical phenomena is usually associated with the experiments of Galvani and Volta around the turn of the nineteenth century. In 1791, Luigi Galvani inadvertently ran a current through a frog's leg and noted the convulsive response. Subsequent experiments with dissimilar metal strips demonstrated the galvanic principle. Although there is circumstantial evidence that copper-iron cylinders made by the Parthians 2000 years ago were primitive batteries, the invention of the battery is usually attributed to Alessandro Volta, who constructed a "pile" from alternate disks of silver and zinc separated by salt-soaked cloth. The connection between chemical and electrical phenomena was confirmed in Volta's experiments and in those of Nicholson and Carlisle, who first electrolyzed water in 1800. Quantitative understanding of the relationships between chemical reaction and electrical charge came in 1830 with Faraday's laws. The concept of electrodeposition was discovered about the same time. A prescient article in the first issue of *Scientific American* in 1845 stated: "This incomprehensible art . . . is truly valuable and must prevail extensively, notwithstanding the disadvantage to which its reputation has been subjected . . ."

Although the fuel cell is commonly associated with space-age technology, its invention is nearly 150 years old. Sir David Grove constructed the first fuel cell from platinum strips immersed in "acidulated water." Grove must also be credited with the first fuel-cell testing program: "A shock was given which could be felt by five persons joining hands, and which taken by a single person was painful." Because of the high cost of hydrogen, the early fuel cell could not compete with batteries, and commercial development was not undertaken. Many novel fuel-cell systems have been subsequently devised, but major development efforts commenced only with impetus from the space program. Fuel cells for terrestrial applications are still in an experimental stage.

Many important processes and electrochemical devices still in use today were conceived in the latter half of the

nineteenth century. Electrochemical routes for producing aluminum and chlorine were devised and soon dominated those industries. The common zinc battery, the dry cell, and the lead-acid battery were all invented in this era.

Serious attempts to quantify the design of electrochemical processes began in the 1920s. The concept of "throwing power" was formulated to characterize the uniformity of an electrodeposit. In the 1940s, methods for simulating the distribution of reaction rates (current distribution) on an electrode surface were described. Several investigators recognized the mathematical similarity between equations describing the current distribution and equations used in fields such as electrostatics, hydrodynamics, and heat conduction. Applicable solutions were subsequently adapted to electrochemical analogs. These early simulations gave approximate solutions for a large class of problems, but effects of electrode kinetics and mass transfer were not rigorously taken into account. The formal synthesis of electrochemistry with engineering principles began in the 1950s and emerged from groups headed by Norbert Ibl in Switzerland and Charles Tobias in the United States. In their early work they devised new techniques for both analysis and measurement of electrochemical phenomena. Effects of hydrodynamics, gas evolution, and electrode geometry were rigorously quantified in generalized design equations. Sophisticated models of electrochemical processes are now available, and the solution of realistic problems is possible through computer simulation.

II. BASIC PRINCIPLES

A. Cell Description

An electrochemical cell consists of two electrodes and an electrolyte through which ions are conducted. The electrodes must be capable of conducting electrons through an external circuit to provide continuity for the charge transfer process. A general cell schematic appears in [Fig. 1](#). In this example, electrical energy is provided to the electrodes. Such a driven device is called an electrolytic cell, whereas an energy-producing device is called a galvanic cell. Under steady-state conditions, chemical species are reduced at one electrode (cathode) and are oxidized at the other electrode (anode). A short-circuited galvanic cell can be considered as a model for corrosion processes. In corroding systems, an electrode (usually a metal) is oxidized, but no useful work is produced. In such systems, oxygen or hydrogen ions are often reduced (at a corresponding rate) on the same surface or on another in electrical contact.

Historically, various sign conventions have been adopted for charge flow, electrode potential, and reaction direction. Benjamin Franklin arbitrarily called the charge

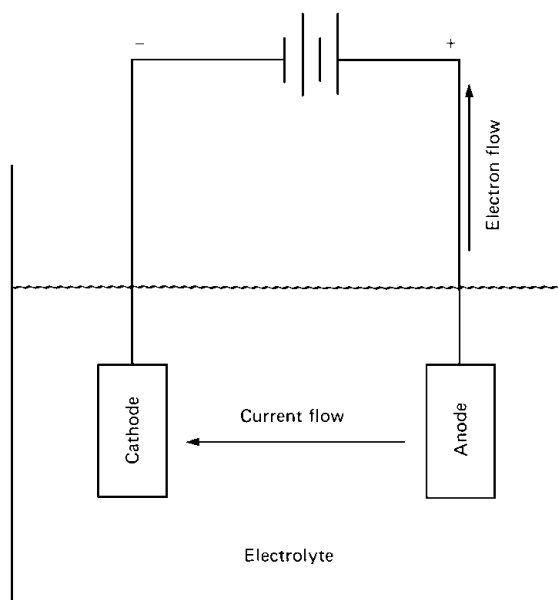


FIGURE 1 Schematic of an electrochemical cell. Electrodes are immersed in electrolyte. The charge is transported by ions in the electrolyte and by electrons in the external circuit.

caused by rubbing glass on silk “positive.” Current flow was originally defined in terms of the flow of positive charges. Although we now recognize that negative electrons carry the charge in a conductor, the original convention is so well-established that its use is still universal.

B. Faraday’s Law

The correspondence between charge flow and chemical reaction was established by Faraday:

$$m = \frac{MIt}{nF}, \quad (1)$$

where m is the mass of the substance produced, M the atomic or molecular weight of the species, I the current, t the time, n the number of electrons participating in the electrode reaction, and F Faraday’s constant (96,500 C). The product of the current and time gives the total charge passed. If the current is not constant, the charge is calculated by integrating the current over the time or is measured with a coulometer.

C. Thermodynamics

For engineering purposes, thermodynamic calculations are useful in several respects. First, they tell us whether a proposed electrochemical system can proceed spontaneously in a given direction. Second, they tell us the maximum work that can be derived from a given cell or, conversely, the minimum work that must be expended

to produce desired products. It is important to recognize that thermodynamic calculations yield information regarding equilibrium states but tell us nothing about the rate at which an equilibrium is attained. Calculation of the rate, which is essential in a design calculation, must be obtained from knowledge of the electrode kinetics and mass-transport limitations.

A large body of thermodynamic data has been amassed over the last century, and it is of obvious value to relate electrochemical variables to these data. One such relation can be developed by recognizing that the maximum work performed by a closed system at constant temperature and pressure is given by the change in Gibbs free energy (ΔG) of the system. In an ideal electrochemical system the change in free energy, which results from chemical reaction, must be equal to the product of the charge and the potential difference through which the charge falls:

$$\Delta G = -nFE, \quad (2)$$

where n is the number of electrons participating in the reaction and E is the reversible cell potential.

From thermodynamic considerations the maximum energy that can be derived from a specified mass of reactants can be calculated. This calculation is of particular interest in the design of portable energy sources. The theoretical specific energy is the ratio of Gibbs free energy of the reaction to the mass of the reactants:

$$\text{Theoretical specific energy} = \frac{\Delta G}{\sum_{\text{reactants}} M_i}. \quad (3)$$

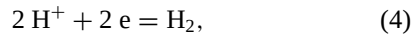
In some calculations the mass of the reactants (especially oxygen derived from the air) that do not need to be transported is not added to the total mass.

D. Potential

The reversible cell potential is the maximum potential that an ideal galvanic device can attain. Because of irreversibilities, the potential difference of a practical galvanic device is always lower. To optimize cell performance, we want to minimize the irreversibilities at a specified current density. A knowledge of the overall cell potential does not give us information regarding the sources of the irreversibilities; detailed knowledge of the individual electrode processes is required for this purpose. To calculate the losses at a particular electrode, we need to know its reversible potential, but this quantity cannot be uniquely specified because there is no absolute zero of potential. As a way of overcoming this difficulty, a specific electrode reaction has been arbitrarily chosen as the standard to which all other electrode systems can be referred. The universal reference electrode is the hydrogen electrode:

TABLE I Standard Electrode Potentials

Electrode reaction	E^0 (V)
$\text{Au}^{3+} + 3 e = \text{Au}$	1.50
$\text{O}_2 + 4 \text{H}^+ + 4 e = 2 \text{H}_2\text{O}$	1.23
$\text{Fe}^{3+} + e = \text{Fe}^{2+}$	0.77
$\text{O}_2 + 2 \text{H}_2\text{O} + 4 e = 4 \text{OH}^-$	0.40
$\text{Cu}^{2+} + 2 e = \text{Cu}$	0.34
$2 \text{H}^+ + 2 e = \text{H}_2$	0.00
$\text{Fe}^{2+} + 2 e = \text{Fe}$	-0.44
$\text{Zn}^{2+} + 2 e = \text{Zn}$	-0.76
$\text{Al}^{3+} + 3 e = \text{Al}$	-1.60



where the hydrogen ions are at unit activity, and the hydrogen gas is at unit fugacity; the reversible potential for this electrode is defined as zero. Several standard electrode potentials are listed in Table I. By convention all electrode reactions are written as reductions.

The theoretical cell potential under standard conditions can be calculated by combining any two reactions of interest. The reversible cell potential is given by subtracting the more negative number from the more positive. The reaction associated with the more positive potential proceeds spontaneously in the direction indicated in Table I. An overall reaction can be indicated by reversing the reaction associated with the more negative potential and multiplying one of the reactions by a constant if the electrons participating in each reaction are not equal.

Operation of an actual electrochemical process invariably takes place under conditions other than those specified for the standard electrode potentials. Since electrode potentials generally vary with temperature, pressure, and concentration, it is necessary to calculate the reversible potential under appropriate conditions. Frequently, the differences are small, and approximate methods are used to calculate the corrections.

The variation in Gibbs-free-energy change with temperature at constant pressure is given by

$$\left(\frac{\partial \Delta G}{\partial T} \right)_P = -\Delta S, \quad (5)$$

where ΔS is the entropy change of the reaction. Combining this with Eq. (2), we obtain

$$\left(\frac{\partial E}{\partial T} \right)_P = \frac{\Delta S}{nF}. \quad (6)$$

As an approximation, the entropy change can be treated as a constant, and the change in reversible potential can be calculated directly.

Variations with pressure are given by

$$\left(\frac{\partial \Delta G}{\partial P} \right)_T = \Delta V \quad (7)$$

or

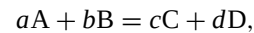
$$\left(\frac{\partial E}{\partial P} \right)_T = -\frac{\Delta V}{nF}, \quad (8)$$

where ΔV is the volume change of the reaction. If ideal behavior can be assumed,

$$\left(\frac{\partial E}{\partial P} \right)_T = -\frac{\Delta NRT}{nFP}, \quad (9)$$

where ΔN is the change in the number of moles of gaseous constituents and R is the gas constant. For condensed phases, pressure corrections are usually neglected.

Concentration corrections can be estimated from the Nernst equation. For a reaction



the Nernst equation is

$$E = E^0 - \frac{RT}{nF} \ln \frac{[C]^c [D]^d}{[A]^a [B]^b}, \quad (10)$$

where the quantities in brackets refer to concentrations. In this equation, activity coefficient corrections and liquid junction potentials are neglected.

E. Reference Electrodes

In principle, we can measure the potential of an electrode with a hydrogen reference electrode. We can also calculate the reversible potential of the cell composed of the electrode of interest and the hydrogen reference electrode. In practice, a hydrogen electrode is difficult to operate properly and is rarely used in engineering measurements. Instead, commercially available reference electrodes (e.g., calomel, Ag/AgCl, and Hg/HgO) are used.

Because of irreversibilities associated with electrode kinetics and concentration variations, the potential of an electrode is different from the equilibrium potential. This departure from equilibrium, known as the overpotential, can be measured with a reference electrode. So that significant overpotential at the reference electrode can be avoided, the reference electrode is usually connected to the working electrode through a high-impedance voltmeter. With this arrangement the reference electrode draws negligible current, and all of the overpotential can be attributed to the working electrode.

F. Ion Conduction

Conduction in electrolytes is due to the movement of positive and negative ions in an electric field. The conductivity is proportional to the density and mobility of charge

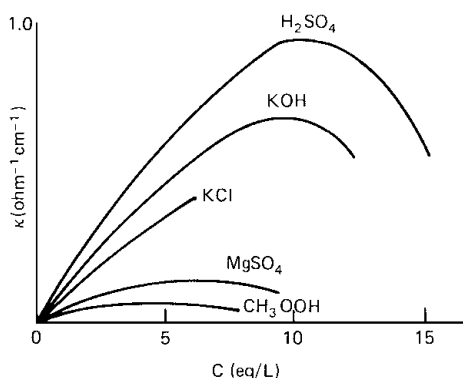


FIGURE 2 Electrolyte conductivity as a function of concentration for common aqueous electrolytes at 25°C.

carriers. Typically, the conductivity of an aqueous electrolyte is between 0.001 and $1 \text{ ohm}^{-1} \text{ cm}^{-1}$. By contrast the electrical conductivity of a metal is of the order of $100,000 \text{ ohm}^{-1} \text{ cm}^{-1}$. Most ionic solutions increase in conductivity with increasing ionic concentration and with increasing temperature. Many solutions exhibit a conductivity maximum that is due to incomplete dissociation of the solute molecules. Salt solutions typically increase in conductivity by about 2% per °C. The conductivities of several common electrolytes are shown in Fig. 2.

It is usually desirable to use high-conductivity electrolyte in an electrochemical process. Ohmic losses, which are inversely proportional to conductivity, result in increased energy consumption. Because the additional energy is converted to heat, low-conductivity electrolytes may require increased thermal management. In industrial practice both the temperature of the electrolyte and the concentration of reacting ions are maintained at relatively high levels. Production of hydrogen by the electrolysis of water is carried out at 85°C in 6 N KOH solution. Another common technique for increasing conductivity is to increase the concentration of charge carriers by adding compounds that dissociate in the solvent but do not participate in the electrode reactions. Such compounds are called supporting or indifferent electrolytes. For instance, adding sufficient sulfuric acid to a copper sulfate solution can increase the conductivity by an order of magnitude. In the electrodeposition of copper, sulfuric acid does not react, but it is frequently added as a supporting electrolyte.

G. Electrode Kinetics

The rate at which an electrochemical process proceeds is governed by the intrinsic electrode kinetics or by mass-transport processes. If reactants are readily available at an electrode surface, then mass-transport limitations do not govern the overall rate; in this section we shall consider this case, in which sluggish kinetics govern the rate.

Many of the concepts from ordinary chemical kinetics have counterparts in electrode kinetics. In both types of systems, energy barriers must be surmounted by the reactants to form products, and increasing the temperature increases the probability that this barrier can be overcome. The important difference in electrochemical systems is that the reaction rate can be increased by increasing the potential difference at the electrode surface. In fact, a significant advantage of electrochemical processes is that an increase in overpotential of 1 V can increase the reaction rate by a factor of 10^8 . In an ordinary chemical reaction a temperature increase of several hundred degrees centigrade is required to produce an equivalent change.

Electrode kinetics are influenced by the potential difference established across a layer immediately adjacent to the electrode surface. As the electrode is polarized, charges build up on the surface of the electrode, and a corresponding charge distribution of opposite sign builds up in the solution about 10 Å from the electrode surface. These two separated regions of charge are referred to as the double layer. The original model for the double layer, proposed by Helmholtz in 1879, was a parallel-plate capacitor. Since the distance between the parallel layers of charge is so small, even a modest potential difference of 100 mV across the double layer leads to an enormous electric field strength, more than 10^6 V/cm . More detailed models of the double layer have subsequently been developed, but the general concept of electrode kinetics being influenced by the strong field adjacent to the electrode surface is still valid.

The passage of a net current through an electrode implies that the electrode is no longer at equilibrium and that a certain amount of overpotential is present at the electrode–electrolyte interface. Since the overpotential represents a loss of energy and a source of heat production, a quantitative model of the relationship between current density and overpotential is required in design calculations. A fundamental model of the current–overpotential relationship would proceed from a detailed knowledge of the electrode reaction mechanism; however, mechanistic studies are complicated even for the simplest reactions. In addition, kinetic measurements are strongly influenced by electrode surface preparation, microstructure, contamination, and other factors. As a consequence, a current–overpotential relation is usually determined experimentally, and the data are often fitted to standard models.

A somewhat general model is that represented by the Butler-Volmer equation,

$$i = i_0 \left[\exp\left(\frac{\alpha_a F}{RT} \eta_s\right) - \exp\left(-\frac{\alpha_c F}{RT} \eta_s\right) \right], \quad (11)$$

where i_0 is the exchange-current density, α_a the anodic transfer coefficient, and α_c the cathodic transfer

coefficient. The exchange-current density and the transfer coefficients can be determined from experimental data. Transfer coefficients typically fall in a range between 0.2 and 2; the exchange-current density varies widely, between 10^{-14} and 10^{-1} A/cm². For copper deposition from aqueous electrolyte near room temperature, $i_0 = 0.001$ A/cm², $\alpha_c = 0.5$, and $\alpha_a = 1.5$, and the Butler-Volmer equation becomes

$$i = 10^{-3}[\exp(58.06 \eta_s) - \exp(-19.35 \eta_s)]. \quad (12)$$

A plot of this relation appears in Fig. 3. Since the transfer coefficients are not equal, the curve is not symmetric about the origin.

Most industrial processes are operated at current densities of more than 50 mA/cm². In this range the overpotential is relatively high, and one of the terms in the Butler-Volmer equation can be neglected. By convention the anodic overpotential is positive, and the cathodic overpotential is negative. If the anodic overpotential is high, then the second term of the Butler-Volmer equation can be neglected:

$$i = i_0 \exp\left(\frac{\alpha_a F}{RT} \eta_s\right), \quad (13)$$

or

$$\eta_s = \frac{RT}{\alpha_a F} \ln \frac{i}{i_0}. \quad (14)$$

Expressed in terms of common logarithms,

$$\eta_s = 2.3 \frac{RT}{\alpha_a F} \log \frac{i}{i_0}. \quad (15)$$

This is the Tafel equation and it is commonly used in design applications. The prelogarithmic term is of the order

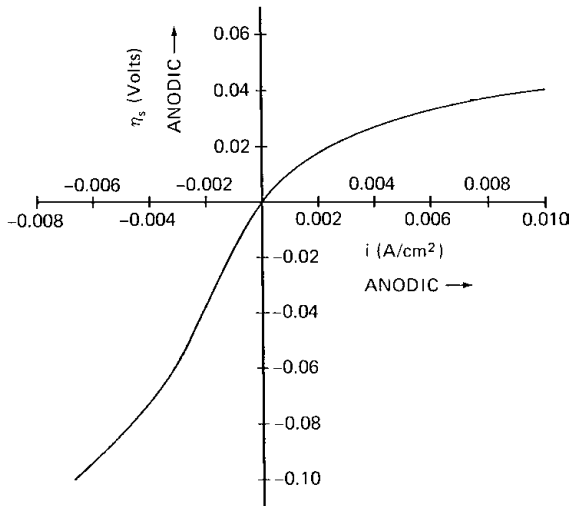


FIGURE 3 Current density-overpotential curve for the Cu/CuSO₄ system at 25°C. The exchange-current density is 0.001 A/cm², $\alpha_a = 1.5$, and $\alpha_c = 0.5$.

TABLE II Approximate Values of Exchange-Current Densities

Reaction	Electrode material	Temperature (°C)	i_0 (A/cm ²)
Hydrogen oxidation	Pt	25	10^{-3}
Hydrogen oxidation	Hg	25	10^{-13}
Oxygen reduction	Pt	25	10^{-10}
Oxygen reduction	Au	25	10^{-12}
Ethylene oxidation	Pt	80	10^{-10}
Copper deposition	Cu	25	10^{-3}

of 100 mV (i.e., the overpotential increases by 100 mV for each factor of 10 increase in the current density).

Less frequently, the exponential terms in the Butler-Volmer equation are small and can be linearized, in which case we obtain

$$i = \frac{(\alpha_a + \alpha_c) i_0 F \eta_s}{RT}. \quad (16)$$

The linear approximation, while not strictly valid at high current densities, is frequently employed as an engineering approximation. This approach is justifiable if the current density variations in a cell are small.

Since the exchange-current density varies over such a wide range, its value is taken as a measure of the sluggishness of reaction kinetics. In this sense an electrode system with a high exchange-current density is considered reversible, and one with a low exchange-current density is irreversible. Typical values are listed in Table II. The central role that the exchange-current density plays in determining surface overpotential is illustrated in Fig. 4. At a current density of 100 mA/cm², the surface

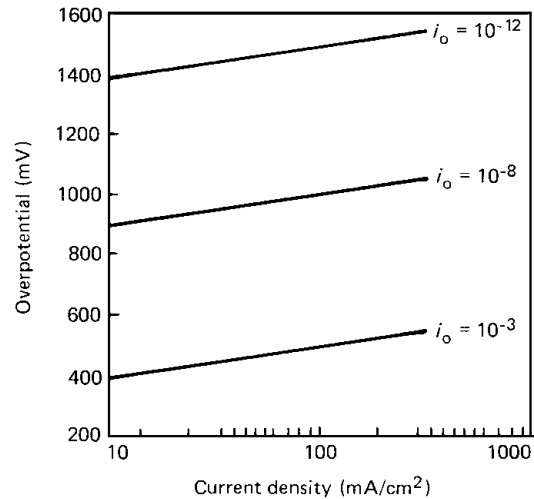


FIGURE 4 Overpotential versus current density when the Tafel slope is 100 mV/decade. Low values of exchange-current density cause significant increases in overpotential at a specified current density.

overpotential is 500 mV when the exchange-current density is 10^{-3} A/cm², whereas the overpotential is 1500 mV for $i_0 = 10^{-12}$ A/cm². This result has especially important consequences for electro-chemical energy conversion devices. From Table II we see that most hydrocarbon oxidations and oxygen reductions are relatively irreversible. In general, the exchange-current density is highest on noble metal surfaces. Because of the irreversibility of these reactions, a practical device for the production of electricity through the direct electrochemical oxidation of hydrocarbons has not been devised.

H. Passivity

The curve shown in Fig. 3 cannot proceed indefinitely in either direction. In the cathodic direction, the deposition of copper ions proceeds from solution until the rate at which the ions are supplied to the electrode becomes limited by mass-transfer processes. In the anodic direction, copper atoms are oxidized to form soluble copper ions. While the supply of copper atoms from the surface is essentially unlimited, the solubility of product salts is finite. Local mass-transport conditions control the supply rate; so a current is reached at which the solution supersaturates, and an insulating salt-film barrier is created. At that point the current drops to a low level; further increase in the potential does not significantly increase the current density. A plot of the current density as a function of the potential is shown in Fig. 5 for the zinc electrode in alkaline electrolyte. The sharp drop in potential is clearly observed at -0.9 V versus the standard hydrogen electrode (SHE). At more positive potentials the current density remains at a low level, and the electrode is said to be passivated.

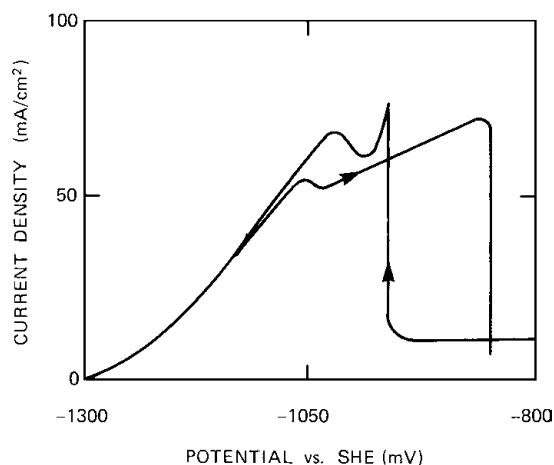


FIGURE 5 Typical potential sweep diagram on a zinc electrode. Current density decreases rapidly near -900 mV versus SHE (standard hydrogen electrode) as reaction products cover the electrode surface and passivate the electrode.

In most corrosion processes passivity is desirable because the rate of electrode dissolution is significantly reduced. The rate of aluminum corrosion in fresh water is relatively low because of the adherent oxide film that forms on the metal surface. A thicker film can be formed on the surface by subjecting it to an anodic current in a process known as anodizing. In most electrochemical conversion processes passive films reduce the reaction rate and are, therefore, undesirable.

III. MASS TRANSPORT

In an electrodeposition process, ions must be transported to the electrode surface and subsequently react by gaining electrons to form metal atoms. Mass transport and electrode kinetics are the individual rate processes that determine the overall deposition rate. Since these rate processes act in series, it is the slowest step that governs the overall rate. For engineering calculations, it is useful to determine the rate-limiting step and to simplify the calculation by neglecting or crudely approximating the remaining steps. If an electrochemical process is limited by mass-transport processes, we must calculate the flux of ions or molecules at the electrode surface. If a gaseous, reactant, such as oxygen, is the limiting species, we can calculate its flux from the ordinary laws governing diffusion and convection; however, for ionic species we must also account for the flux due to the influence of the electric field on the charged species.

A. Governing Equations

A mathematical description of an electrochemical system should take into account species fluxes, material conservation, current flow, electroneutrality, hydrodynamic conditions, and electrode kinetics. While rigorous equations governing the system can frequently be identified, the simultaneous solution of all the equations is not generally feasible. To obtain a solution to the governing equations, we must make a number of approximations. In the previous section we considered the mathematical description of electrode kinetics. In this section we shall assume that the system is mass-transport limited and that electrode kinetics can be ignored.

Species flux can be described by the Nernst-Planck equation,

$$N_i = -z_i u_i F c_i \nabla \phi - D_i \nabla c_i + c_i v, \quad (17)$$

where N_i is the flux of species i , z the charge on the ion, u_i the mobility, c_i the concentration of i , $\nabla \phi$ the potential gradient, D_i the diffusivity of i , and v the bulk velocity. The first term on the right represents the flux due to

migration, the movement of charged species under the influence of an electric field. The second term is the flux due to diffusion, and the third term is the flux due to convection. This expression is strictly correct for extremely dilute solutions; however, it is generally applied to more concentrated solutions and used as a reasonable engineering approximation.

The current is due to the motion of charged species:

$$i = F \sum_i z_i N_i. \quad (18)$$

At steady state the net input of a reacting species in the electrolyte is zero. If we assume that reactions occur only at the electrode surface, then the material balance can be expressed as

$$\nabla \cdot N_i = 0. \quad (19)$$

Because the electrical forces between charged species are so large, the positive and negative particles have a strong tendency to associate. On a macroscopic level, charge separation cannot be detected in the bulk electrolyte, and the solution is electrically neutral:

$$\sum_i z_i c_i = 0. \quad (20)$$

These four equations form the basis for a description of the mass transport in electrolytic solutions. To solve these equations, we must calculate the bulk solution velocity from a knowledge of the fluid mechanics.

Solutions to this system of equations depend on the cell geometry and on the boundary conditions; therefore, generally valid solutions cannot be obtained. With certain simplifying assumptions, the equations reduce to familiar forms, and solutions can be obtained for large classes of problems.

If temperature and concentration variations are neglected, then an expression for the potential distribution in the bulk electrolyte is given by Laplace's equation,

$$\nabla^2 \phi = 0. \quad (21)$$

If we neglect the overpotential at the electrodes, then the boundary conditions for solving this problem are the constant electrode potentials. This type of problem has exact analogs in electrostatics, and many generalized solutions for symmetric configurations are available. In this type of problem, the current density is proportional to the potential gradient, and the current distribution can be calculated from Ohm's law:

$$i = -\kappa \nabla \phi. \quad (22)$$

For the solution of a salt composed of two ionizable species (binary electrolyte), the four basic equations can be combined to yield the convective diffusion equation for steady-state systems:

$$v \cdot \nabla c = D \nabla^2 c, \quad (23)$$

where

$$D = \frac{z_+ u_+ D_- - z_- u_- D_+}{z_+ u_+ - z_- u_-}. \quad (24)$$

The convective diffusion equation is analogous to equations commonly used in dealing with heat and mass transfer. Similarly, if migration can be neglected in a multicomponent solution, then the convective diffusion equation can be applied to each species,

$$v \cdot \nabla c_i = D_i \nabla^2 c_i. \quad (25)$$

The hydrodynamic conditions influence the concentration distribution explicitly through the velocity term present in the convective diffusion equation. For certain well-defined systems the fluid flow equations have been solved, but for many systems, especially those with turbulent flow, explicit solutions have not been obtained. Consequently, approximate techniques must frequently be used in treating mass transfer.

B. Mass-Transfer Boundary Layer

Consider the process of plating copper on a plane electrode. Near the electrode, copper ions are being discharged on the surface and their concentration decreases near the surface. At some point away from the electrode, the copper ion concentration reaches its bulk level, and we obtain a picture of the copper ion concentration distribution, shown in Fig. 6. The actual concentration profile resembles the curved line, but to simplify computations, we assume that the concentration profile is linear, as indicated by the dashed line. The distance from the electrode where the extrapolated initial slope meets the bulk concentration line is called the Nernst diffusion-layer thickness δ . For order of magnitude estimates, δ is approximately 0.05 cm in unstirred aqueous solution and 0.01 cm in lightly stirred solution.

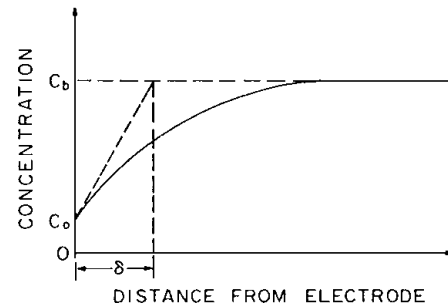


FIGURE 6 Nernst diffusion-layer model. The solid line represents the actual concentration profile, and the dashed line for c_0 the Nernst model concentration profile.

It is clear from Fig. 6 that the concentration of a reacting species decreases at the electrode surface as the current is increased. The minimum concentration is zero at the surface, which corresponds to the maximum rate at which the electrodeposition reaction can proceed. The current density corresponding to this maximum rate is called the limiting current density i_l , which can be approximated by

$$i_l = \frac{nFD_i c_b}{\delta}, \quad (26)$$

where c_b is the bulk concentration of copper ions. Processes occurring at the limiting current represent the case in which only the mass-transfer limitations must be considered, and the kinetic limitations and ohmic effects can be neglected. Because there are numerous correlations for the limiting current density, many cases of engineering interest can be treated in an approximate manner.

C. Concentration Overpotential

The concentration of reacting species can vary significantly across the relatively thin mass-transfer boundary layer. When the reacting species are ions, a potential difference, called the concentration overpotential, arises because of these gradients. When operation is occurring at less than 90% of the limiting current density, the magnitude of the concentration overpotential is relatively small (of the order of 10 mV). Approximate expressions for estimating concentration overpotential are available for binary electrolyte and for the case in which supporting electrolyte is present. In the latter case the expression for concentration overpotential is

$$\eta_c = \frac{RT}{nF} \ln \left(1 - \frac{i}{i_l} \right). \quad (27)$$

IV. CURRENT DISTRIBUTION

A. Classification

Overall power requirements for an electrolytic process are determined from a knowledge of the total current and the applied potential; however, more detailed knowledge of the distribution of reaction rates (current distribution) is required in an optimization of system performance. Although local current densities can usually be measured, it is always desirable to develop a mathematical model of the process and to simulate the effects of changes in operating conditions.

In making a calculation of the current distribution, we need to select only the important variables for use in the simulation. If the geometry is symmetric and only one or

two variables control the reaction rate, there is a possibility that the current distribution has been calculated over a range of operating variables, and we can use these solutions directly. If this is not the case, then it is likely that a model must be constructed, and computer techniques are required in the solution.

Current distribution problems are usually classified according to the rate-limiting process:

1. Primary current distribution. The current distribution is governed solely by the electric field. No other effects are considered.
2. Secondary current distribution. Both field effects and the effects of sluggish reaction kinetics are considered.
3. Tertiary current distribution. Field effects, kinetic limitations, and mass-transfer limitations are all considered.

The complexity of a model increases as we proceed from the primary to the tertiary distribution and as the number of spatial dimensions that are considered increases. Essentially all published solutions have been reduced to one or two dimensions, and most include only simulations of the primary and secondary current distributions. For the special case in which only mass transport is limiting, a large number of correlations for the current distribution are available.

B. Primary Current Distribution

The primary current distribution represents the distribution resulting solely from resistance to current flow in the electrolyte. Since temperature and concentration variations as well as overpotential are neglected, this type of current distribution is usually easy to calculate.

Laplace's equation governs the potential distribution [Eq. (21)]. Since overpotential is ignored, the potential immediately adjacent to the electrodes is constant. At insulated surfaces the normal potential gradient must be zero. These two requirements dictate the boundary conditions for the differential equation.

Models for phenomena such as heat conduction, fluid flow, and diffusional mass transfer are also based on Laplace's equation. Consequently, many solutions to the potential distribution problems or the analogous problems in other fields are available. The current distribution can be obtained from the potential distribution through Ohm's law [Eq. (22)].

If the assumptions inherent in the primary current distribution model are reasonable for the system being considered, then a simulation of the system behavior is relatively straightforward.

C. Secondary Current Distribution

The secondary current distribution is calculated by including the effects of the ohmic drop in the electrolyte and the effects of sluggish electrode kinetics. While the secondary distribution may be a more realistic approximation, its calculation is more difficult; therefore, we need to assess the relative importance of electrode kinetics to determine whether we can neglect them in a simulation.

Kinetic limitations are manifested by surface overpotential. A plot of surface overpotential on the ordinate versus current density on the abscissa can be used to determine a so-called polarization resistance. If the slope of the line is relatively steep, then small changes in the current density give rise to large changes in the overpotential; this implies that the electrode reaction is sluggish, and the polarization resistance ($\partial\eta_s/\partial i$) is large. Conversely, a relatively flat line is characteristic of a reaction with low polarization resistance. A dimensionless parameter, called the Wagner number, characterizes the ratio of the polarization resistance to the electrolyte resistance:

$$Wa = \kappa \frac{\partial\eta_s}{\partial i} \bigg|_{i_{avg}} / L, \quad (28)$$

where L is a characteristic dimension of the system. As Wa approaches zero, the kinetic limitations are negligible, and the primary current distribution is appropriate. A Wagner number equal to one indicates that the effects of kinetics and electrolyte resistance are both significant, and a secondary current distribution model is appropriate. As Wa becomes very large, the effects of electrolyte resistance are both significant, and a secondary current distribution model is appropriate. As Wa becomes very large, the effects of electrolyte resistance can be neglected, and the current distribution becomes more uniform.

Consider the wavy electrode and the planar counterelectrode shown in Fig. 7. The resistance to ion flow is depicted

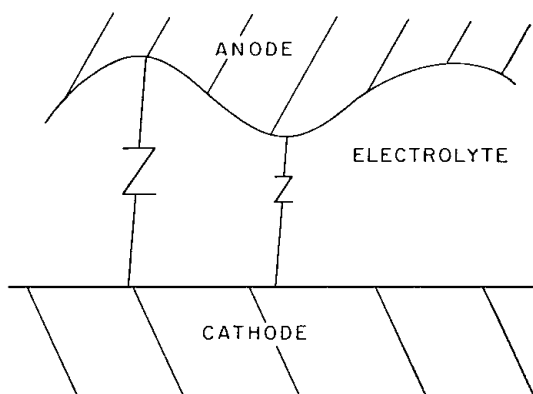


FIGURE 7 Primary current distribution on a wavy electrode. Resistance to current flow is represented schematically by the size of the resistors.

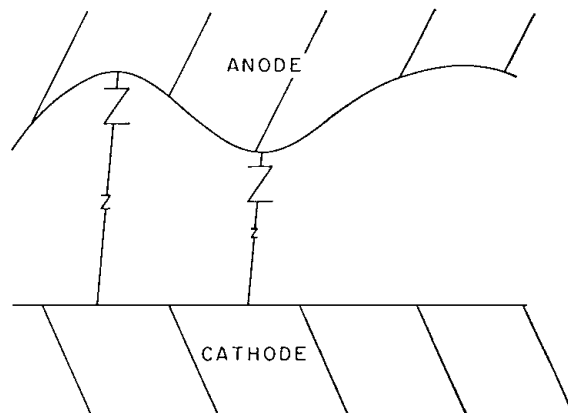


FIGURE 8 Secondary current distribution. When surface overpotential governs the current distribution, small differences in solution resistance (represented by smaller resistors) can be neglected, and the current distribution becomes more uniform.

schematically by resistors whose sizes are proportional to their magnitudes. If the electrolyte is resistive (κ is low), then that resistance dominates, and the primary current distribution model is appropriate. The relative amount of current reaching any portion of the wavy electrode is related to its distance from the counterelectrode. This is indicated on the figure by a larger resistor for the longer distance. The current density is nonuniform because the point closest to the counterelectrode has a higher current density than that in the depression.

By contrast, consider the same system in which kinetic limitations dominate. Ions must be transported through the solution and then react at the electrode surface. These processes can be modeled as resistances in series, as shown in Fig. 8. In this case the larger resistors represent the polarization resistance, and the small differences in electrolyte resistance make little differences in the current density reaching any portion of the electrode. Consequently, the current distribution is relatively uniform. For this particular geometry the amplitude of the electrode is a good choice for the characteristic dimension. At a specified average current density, increasing the amplitude reduces Wa . As intuitively expected, a smaller value for the Wagner number implies that the current distribution becomes more nonuniform.

D. Tertiary Current Distribution

At an appreciable fraction of the limiting current, it is usually not justified to neglect concentration variations—and resulting overpotential—near the electrode. In a general model we need to consider the electric field, kinetic limitations, and concentration variations. The problem is rendered more difficult by the need to know the system hydrodynamics, which, in turn, influence the concentration

distributions. For a few systems, important in electrochemical applications, the detailed fluid behavior is known. Even with this knowledge, finding a solution to the current distribution problem for all but the simplest geometries is a formidable task. The hydrodynamic conditions for laminar flow at a rotating disk and between plane parallel electrodes have been quantitatively described. These are among the few systems for which fairly rigorous tertiary current distributions have been obtained.

When a system is operating at the limiting current, rather than at an appreciable fraction of the limiting current, the problem is very much simplified. Such problems can be classified as mass-transport limited. Usually, the limiting current density is correlated with dimensionless numbers. Most forced-convection correlations take the form

$$\text{Sh} = f(\text{Re}, \text{Sc}), \quad (29)$$

where Sh (Sherwood number) is related to the limiting current density, Re (Reynolds number) characterizes the hydrodynamics, and Sc (Schmidt number) is related to transport properties of the fluid. Both laminar and turbulent flow problems are treated over a wide range of operating and physical parameters in this manner.

E. Current Distribution Characteristics

Several cell configurations are common in electrochemical research and in industrial practice. The rotating disk electrode is frequently used in electrode kinetics and in mass-transport studies. A cell with plane parallel electrodes imbedded in insulating walls is a configuration used in research as well as in chemical synthesis. These are two examples of cells for which the current and potential distributions have been calculated over a wide range of operating parameters. Many of the principles governing current distribution are illustrated by these model systems.

The rotating disk electrode appears in Fig. 9. It consists of a cylindrical electrode imbedded in an insulating disk.

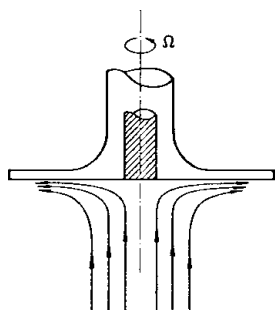


FIGURE 9 Rotating disk electrode. Fluid is drawn uniformly to the electrode surface, and the reactant concentration depends only on the normal distance from the electrode.

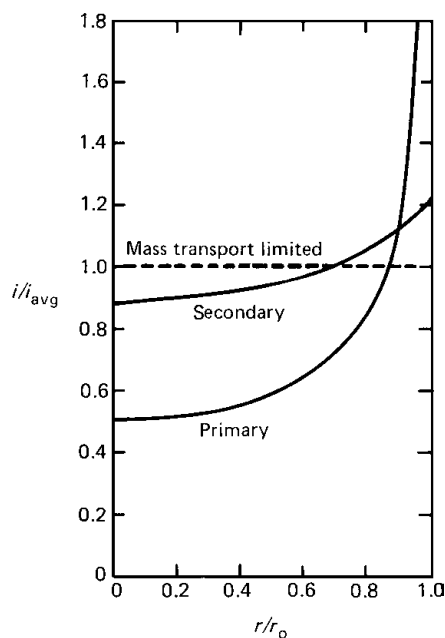


FIGURE 10 Current distribution on a disk electrode. The primary current distribution approaches infinity at the junction of the electrode and the coplanar insulator. The secondary current distribution is more uniform. Average current density is i_{avg} and the electrode radius r_0 .

As the disk spins, it pumps fluid to the surface. For laminar flow, analytical solutions describing the fluid motions have been obtained. In modeling the system the disk is assumed to be immersed in a large volume of electrolyte with the counterelectrode far away.

The primary potential distribution is, by definition, uniform adjacent to the electrode surface, but the current distribution is highly nonuniform (Fig. 10). It is a general characteristic of the primary current distribution that the current density is infinite at the intersection of an electrode and a coplanar insulator. This condition obtains at the periphery of the disk electrode, and the current density becomes infinite at that point. Additional resistance due to kinetic limitations invariably reduces the nonuniformity of the current distribution. In this system the current distribution becomes more uniform as the Wagner number increases. Theoretically, the current distribution is totally uniform as the Wagner number approaches infinity.

In general, the effects of mass-transport limitations are not as easy to characterize. The direction of fluid flow, the flow regime, and the local fluid velocity all influence the current distribution. Fluid flow to the rotating disk is unusual in that fluid velocity normal to the disk is dependent only on the normal distance from the disk surface, and not on radial distance. Because the disk surface is uniformly accessible to incoming reactants, mass-transport limitations tend to reduce the current density in regions of high

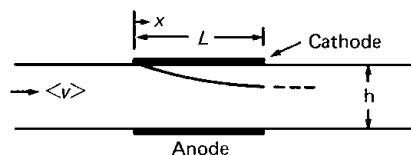


FIGURE 11 Plane parallel electrodes imbedded in insulating walls. Fluid flows from right to left, and the reactant concentration tends to decrease in the direction of fluid flow as the electrochemical reaction progresses.

current density. Therefore, in this system mass-transport limitations cause the current distribution to be more uniform. When the system is operating at the limiting current, the current distribution is completely uniform.

Channel flow between plane parallel electrodes is shown in Fig. 11. This geometry is similar to that of the disk in that an electrode and an insulator intersect in the same plane. Because of many geometric similarities, the general characteristics of the primary and secondary current distributions are similar. At the edges the local current density is infinite for the primary current distribution (Fig. 12). Increasing the kinetic limitations tends to even out the current distribution. The significant contrasts appear in a comparison of the tertiary current distributions. In channel flow, the fluid flows across the electrode rather than normal to it. Consequently, the electrode is no

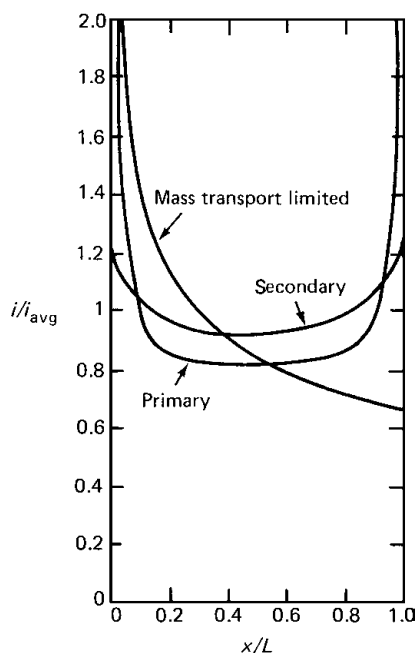


FIGURE 12 Current distribution on plane parallel electrodes. Primary and secondary current distributions are symmetric about a centerline plane. When the reactant concentration is considered, an asymmetric current distribution results.

longer uniformly accessible to reactants. Instead, the reactant concentration is highest at the leading edge and falls as the reaction proceeds farther down the electrode. The resulting current distribution tends to be skewed toward the leading edge.

V. SYSTEM DESIGN

A. Process Modeling

The modeling of electrochemical processes has evolved over the past 50 years to the point where complex problems involving multiple reactions, temperature variations, and physical property variations can be treated. Essentially all contemporary models require iterative computer techniques to simulate system behavior.

Several general techniques are used in the modeling of electrochemical systems. A method for reducing the geometry to its basic configuration is called sectioning. In potential theory problems (primary and secondary current distributions) planes of symmetry can be replaced by insulators. In the channel electrode model it is clear that the plane of symmetry cuts the electrodes through their midpoints. This plane could be replaced by an insulator across which no current flows. The plane surface establishes the boundary condition $\nabla\phi = 0$. As we intuitively expect, the primary and secondary current distributions (Fig. 12) are symmetric about midplane. The same type of procedure can be applied to the infinite sinusoidal wave shown in Fig. 7. The current distribution is symmetric about a properly selected half-wavelength.

Because the kinetic and mass-transport phenomena occur in a thin region adjacent to the electrode surface, this area is treated separately from the bulk solution region. Since kinetic effects are manifested within 100 Å of the electrode surface, the resulting overpotential is invariably incorporated in the boundary conditions of the problem. Mass transport in the boundary layer is often treated by a separate solution of the convective diffusion equation in this region. Continuity of the current can then be imposed as a matching condition between the boundary layer solution and the solution in the bulk electrolyte. Frequently, Laplace's equation can be used to describe the potential distribution in the bulk electrolyte and provide the basis for determining the current distribution in the bulk electrolyte.

While it is usually possible to write the governing equations, effecting a solution can pose many difficulties. Many analytical solutions for symmetric geometries with straightforward boundary conditions have already been solved. It is, therefore, highly unlikely that an analytical solution will be obtainable for novel systems, and some numerical method must be used.

To date, most simulations have been based on the finite difference technique or the finite element method. In both methods the domain of interest is divided into smaller subdomains. Trial solutions for one of the variables are assumed, and these are corrected through continued iteration. Convergence is assumed when the solutions do not change significantly between iterations. While convergence for secondary and tertiary current distribution problems is not ensured, general techniques for promoting convergence are available.

In most cases the accuracy of the solution increases as the domain of interest is more finely divided; however, the computer calculation time also increases with the finer division. An advantage of using finite element and finite difference techniques is that commercial routines are available to solve some of the pertinent equations.

Other methods such as orthogonal collocation and boundary element techniques have also been used. The relative advantages of using the various methods usually involve trade-offs among factors such as programming ease, accuracy of solution, storage capability of the computer, and availability of software.

B. Technical Factors

Effective system design depends on the proper application of the principles of thermodynamics, kinetics, and transport phenomena. Reliable design data are invariably obtained empirically because ab initio computation of design parameters, such as kinetic quantities, are not sufficiently reliable for engineering purposes.

From the basic principles we can make preliminary design estimates. Inefficiencies in a system arise because of voltage losses and because all of the current does not enter into the desired reactions. The minimum potential required to perform an electrolytic reaction is given by the reversible cell potential, a thermodynamic quantity. Additional voltage that must be applied at the electrodes represents a loss that is manifested in a higher energy requirement. The main causes of voltage loss are ohmic drops and overpotentials. The applied potential is equal to the sum of the losses plus the thermodynamic requirement:

$$V_{\text{applied}} = E + \Delta V_{\text{ohm}} + \sum_i \eta_{si} + \sum_i \eta_{ci} \quad (30)$$

Ohmic losses can result from a variety of causes: resistance to ion flow in the electrolyte, resistance in the bus bars, and resistance in membranes used to separate anode and cathode electrolytes. The magnitude of the resistances may change with time as films build up on electrode surfaces or as membranes become contaminated. Surface overpotentials can be estimated from rate expressions such as the Tafel equation, or they can be evaluated from em-

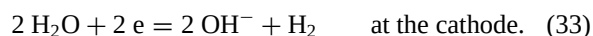
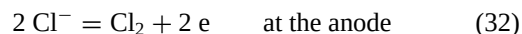
pirical data. If operation significantly below the limiting current is anticipated, concentration overpotentials can be neglected; at higher current densities concentration overpotential estimates are obtainable from Eq. (27).

Current efficiency is defined as the fraction of the total current participating in the desired reaction. The portion of the current that produces undesired products is usually a function of the current density; generally, parasitic reactions are more likely to be favored at higher current densities. Overall energy efficiency is the product of the voltage efficiency times the current efficiency. In an optimization it is useful to examine the magnitudes of the losses from the various sources and to determine whether the major losses can be minimized.

Consider the production of chlorine and caustic soda from brine. This process is one of the most important commercial, electrolytic syntheses; worldwide production of chlorine is currently 30 million tons per year. In one electrochemical route the overall reaction is



and the electrode reactions are



A modern cell operates at approximately 300 mA/cm² and 85°C; the anode electrolyte (anolyte) is 15% NaCl, and the cathode electrolyte (catholyte) is 30% NaOH. A breakdown of the thermodynamic potential and the voltage losses appears in Table III.

It is clear that most of the applied voltage is required for the decomposition process. The voltage efficiency is 2.2 V/3.7 V = 0.6, (i.e., 60% of the potential drop is required to carry out the reaction). The remaining 40% of the potential is converted into heat by various irreversible processes. When this reaction is carried out in a modern cell, the current efficiency is quite high, usually more than 95%. Overall energy efficiency for this process is just under 60%.

From this simple analysis the distribution of energy losses is immediately apparent. While there are several general techniques that can be used to increase efficiency,

TABLE III Components of the Applied Potential in a Chlor-Alkali Cell

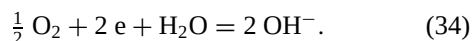
Thermodynamic potential	2.2 V
Anodic overpotential	0.1
Ohmic losses	0.6
Membrane loss	0.5
Cathodic overpotential	0.3
Terminal voltage	3.7 V

their implementation is usually not so straightforward. Various compromises are made to minimize overall cost per unit of product.

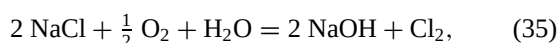
To reduce ohmic loss, one usually has two choices: reduce the electrode separation or increase the electrolyte conductivity. For chlorine production, cells in which both the anode and the cathode contact the membrane have been designed. Such zero-gap cells are expected to replace present designs having cell gaps of several millimeters. Electrolyte conductivity can be increased by raising the temperature and by increasing the concentration of charge carriers. The maximum temperature in aqueous systems is dictated by the boiling point of the medium; and even at lower temperatures, materials problems and corrosion may impose limits. Electrolyte concentration is usually maintained at a relatively high level, and supporting electrolyte is frequently added to increase the conductivity.

Techniques for increasing the reaction rate in electrochemical systems are analogous to those used for ordinary chemical reactions. Increased temperature and catalysis are usually effective. In chlorine production, Raney nickel has been shown to reduce the overpotential by 0.2 V at the cathode. Although chlorine was formerly evolved on graphite anodes, these have been largely replaced by anodes composed of titanium and ruthenium oxide, with a voltage savings of 300 mV. Although higher temperature is advantageous for reducing ohmic losses and surface overpotential, corrosion, phase change, and adverse selectivity ratios must all be considered.

Reducing the thermodynamic requirement is usually most difficult to effect. In some cases modest reductions in reversible potential can be accomplished by changing the temperature or the pressure of the system. Major changes in the thermodynamic requirement are usually possible only by altering the overall reaction. For chlorine production, oxygen reduction has been suggested as an alternate cathode reaction:



The overall reaction then becomes



and the reversible potential is 1.1 V instead of 2.2 V. The primary sacrifice with this route is the loss of hydrogen; however, the hydrogen is of relatively little value because it is usually burned as a fuel. A practical drawback of this scheme is that oxygen reduction is a sluggish process, and the overpotential at this electrode can be significant.

C. Economic Factors

The economic optimum for an electrochemical process usually reflects a compromise between capital costs and

operating expenses. Lower capital costs are incurred in a system in which the electrode surfaces are relatively small and the current density is relatively high (for a specified production rate); however, significant irreversibilities accompany a higher current density, and the energy costs increase. The opposite is true for larger electrode areas: Operating costs are reduced at the expense of capital costs. For low-priced commodity chemicals such as hydrogen and chlorine, the minimum current density must be relatively high (several hundred mA/cm²) to restrict capital costs. The optimum is sensitive to energy costs. Recent rises in electrical costs have put more of a premium on reduced energy consumption through more efficient design. In chlor-alkali cells, energy requirements have dropped from 3500 kWh/metric ton in 1980 to 2800 kWh/metric ton in 1983, and cells under development are approaching 2100 kWh/metric ton.

D. Energy Conversion Systems

Electrochemical devices are being developed for large-scale energy conversion and storage applications. Fuel-cell demonstration units with 4.8-MW outputs are currently being tested. These devices have the advantage of performing a direct conversion from fuel to electricity, thus avoiding Carnot cycle losses. Despite advantages in thermodynamic efficiency, the reliability and overall efficiency are not sufficiently high to displace current thermal-cycle technology. One source of inefficiency stems from the inability of fuel cells to use hydrocarbons directly. The irreversibility associated with using available hydrocarbons, such as ethylene, is a severe limitation (see Table II); moreover, oxygen reduction is also a difficult process to catalyze. Most fuel-cell systems currently under development require hydrogen at the anode, as the electrode kinetics are much more favorable. Conversion of common fuels to hydrogen requires a processing step, which lowers the overall efficiency.

Large-scale energy storage is being considered for electric utility load leveling. In this scheme electrical energy produced during off-peak hours is stored in a secondary (rechargeable) battery and is released back into the grid during peak-demand periods. The main advantage of this mode of operation is that additional capital expenditures, required for peak-load generation equipment, can be avoided. For commercial adoption the economics of the storage system must be advantageous. Currently, the cycle life of most systems is inadequate. A commercial system would need to be capable of a minimum of 2500 cycles or about 10 yr of continuous service. The lead-acid battery can meet this goal, but capital costs for that system are too high to compete with conventional load-following technology.

Electrochemical devices have many advantages that make them attractive for transportation applications. Most electrochemical power sources are pollution-free, quiet, and efficient. These attributes, especially efficiency, have made fuel cells ideal electrical power sources for manned spacecraft. Urban transportation is a large-scale application in which similar attributes are desirable. For stationary systems, device weight is not an important consideration. By contrast, energy per unit weight (specific energy) and power per unit weight (specific power) are of prime importance in the design of systems for transportation uses.

If the specific energy is too low, the battery weight becomes prohibitive. Low specific power implies that vehicle acceleration may be unacceptable. For essentially all systems under consideration, the theoretical specific energy is significantly higher than the minimum requirement of approximately 100 Wh/kg (Table IV). However, because the battery is not totally discharged during each cycle and because a support system (casings, pumps, etc.) is required, actual specific energy is roughly 20% of the theoretical value. The Ragone plot (Fig. 13) shows that most ambient temperature batteries do not meet the minimum specific energy and power requirements (100 W/kg). Power limitations can usually be overcome by higher-temperature operation. Several molten salt systems, operating at 300–700°C, meet these requirements, but materials problems must be overcome before such systems can be used commercially.

E. Future Developments

With the widespread use of laptop computers, cellular telephones, and other portable electrical devices, the need for high energy density power sources has increased. In the past decade, two systems for these purposes have been commercialized: nickel–metal hydride and lithium-ion batteries. For automotive applications, the interest in

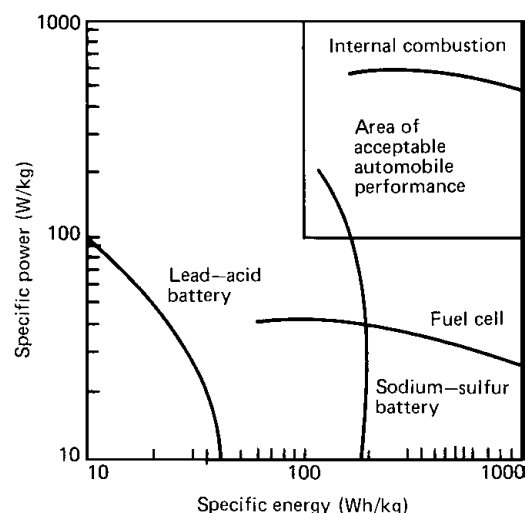


FIGURE 13 Ragone plot. Acceptable automobile performance requires the specific power and specific energy shown in the upper right corner of the plot. Several secondary battery systems can meet these technical objectives.

economical electrical power sources with acceptable reliability, lifetime, and performance has been enhanced by regulations to reduce urban pollution; fuel cells are receiving increased attention for this purpose. In particular, proton exchange membrane (PEM) fuel cells are being developed for automotive applications.

For high-performance applications, lithium-based systems are being developed. Lithium has several potential advantages for battery applications, including low equivalent weight, a highly negative standard potential, and a moderate material cost. When it is coupled with a sulfur cathode, the theoretical specific energy is more than 2300 Wh/kg, among the highest of any couple being considered for commercial development. Because of safety and other practical considerations, however, lithium is often alloyed and other less active cathode materials are used; consequently, the theoretical specific energy of the current generation of lithium-based systems is approximately 500 Wh/kg.

Lithium primary batteries have been standard commercial products for several decades, but a rechargeable version became available only as recently as 1991. Failure of secondary batteries through dendrite formation posed safety problems. The solution to this problem was the development of an innovative design in which lithium ions move between intercalation electrodes. Ions move away from the anode during discharge and reverse the process during charge in what is known as a “rocking-chair” mechanism. In such electrodes the lithium ions occupy interstitial spaces in the host material. During discharge, the lithium ions move from a graphitic carbon anode through

TABLE IV Theoretical Specific Energy for Systems Being Considered in Transportation Applications

Reaction	Theoretical specific energy (Wh/kg)
$\text{Pb} + \text{PbO}_2 + 2 \text{H}_2\text{SO}_4 = 2 \text{PbSO}_4 + 2 \text{H}_2\text{O}$	175
$\text{Zn} + 2 \text{NiOOH} + \text{H}_2\text{O} = \text{ZnO} + 2 \text{Ni(OH)}_2$	326
$2 \text{LiAl} + \text{FeS} = \text{Li}_2\text{S} + \text{Fe} + 2 \text{Al} \quad (T = 450^\circ\text{C})$	458
$2 \text{Na} + 2 \text{S} = \text{Na}_2\text{S}_3 \quad (T = 350^\circ\text{C})$	758
$\text{Zn} + \frac{1}{2} \text{O}_2 = \text{ZnO}$	1360 ^a

^aOxygen is obtained from the air and is not included in the calculation of the reactant mass.

an organic solvent or a polymeric electrolyte to an oxide or a sulfide cathode. Currently, the standard cathode is cobalt oxide; however, considerable research is under way to replace this material, which is toxic and expensive. Lithium batteries based on this intercalation mechanism are called lithium-ion batteries. A good safety record coupled with high energy density has made lithium-ion batteries popular for portable computers, CD players, desktop computer backup, and cellular phones.

The use of a pure lithium anode can potentially be used to produce a battery with a higher energy density than that of the lithium-ion battery. One concept is to fabricate a battery consisting of a lithium foil anode, a polymer electrolyte, and an active sulfur composite cathode. This type of secondary battery, currently in the development stage, would significantly decrease the weight of portable devices. One disadvantage of these lithium-polymer batteries is that the polymer must be operated at elevated temperatures to perform adequately. As mentioned previously, a general problem with lithium secondary batteries has been dendritic growth, which leads to shorting; therefore, potential safety problems associated with the failure of this type of high energy density battery must be addressed. Most lithium designs also require more precise charge control because of their low tolerance for overcharging.

Battery deficiencies have been the major factor impeding the development of commercial electric vehicles. Both batteries and fuel cells have been used in prototypical electric vehicle designs, but factors such as low energy density, high cost, and low cycle life have made commercialization impractical. Only small-scale trials have been conducted to test public acceptance of electric vehicles. Most test vehicles have used lead-acid batteries, which are unlikely to gain general acceptance because their low energy density results in a limited range. Current fuel cells operate most efficiently on hydrogen, which is difficult to store. Hydrogen can be produced from conventional liquid fuels through reforming, but this step requires more processing and added weight.

Internal combustion technology has inherent advantages over battery technology in terms of specific energy and the rate at which energy can be transferred to a vehicle from an external source. The energy content of gasoline is approximately 12,000 Wh/kg, whereas the most energetic battery under development is projected to have a specific energy of 200 Wh/kg. Even with Carnot losses and other inefficiencies, the internal combustion vehicle readily achieves a specific energy on the order of 1000 Wh/kg. Because of the relatively sluggish kinetics of most battery systems, the rate of recharging is slow. A gasoline-powered vehicle can be refueled at a rate roughly equivalent to 100 miles/min, whereas the rate for a battery system is about one or two orders of magnitude slower. With these

factors in mind, researchers are investigating several approaches to incorporate electrochemical technology in to vehicles.

Because of the current limitations of electrochemical power sources for vehicles, several hybrid concepts have emerged. One vehicle is being marketed with a small internal combustion engine coupled with a battery that can deliver and accept charge at high rates for short periods. The engine can be activated when high power is required or when the battery is recharging. In urban driving, the battery will permit operation in an environmentally benign mode.

For the hybrid application, a nickel-metal hydride battery is often used. These batteries have a commercial base in consumer applications, and they have a 50% higher energy density than that of lead-acid batteries. Hydrogen is stored in metal hydride anodes, which are catalytic alloys of metals such as vanadium, titanium, zirconium, and nickel. During discharge the hydrogen is oxidized at the negative electrode and nickel is reduced at the positive electrode. These reactions are fully reversible, and side reactions are minimal; consequently, the battery has a long cycle life.

Interest in fuel cells for transportation is growing rapidly. Operational fuel cells were first demonstrated in the space program beginning with the Gemini and Apollo spacecraft in the 1960s. The low power density and high cost made these configurations impractical for more general applications. The PEM fuel cell is now being considered for use in electric vehicles. This system consists of two porous carbon electrodes separated by an ion-conducting polymer electrolyte, which conducts protons but is impermeable to gas. Catalysts are integrated between the electrodes and the membrane. The anode is supplied with hydrogen and the cathode with air. However, before these systems see widespread application, issues of cost and hydrogen storage must be addressed. Furthermore, the polymer membranes are currently expensive, as are the noble metal catalysts.

All current fuel-cell systems operate most efficiently on hydrogen, but storing this fuel for mobile applications requires a separate, cumbersome system. Another concept is to generate the hydrogen on-site by using the well-established technology of reforming from a liquid fuel, such as methanol or gasoline. Steam reforming of methanol is technically simpler than the partial oxidation of gasoline; however, the existing distribution infrastructure favors hydrocarbon use. An alternative is to use a liquid fuel, which would be more convenient and more compatible with the existing infrastructure. For this purpose the direct methanol fuel cell (DMFC) is the leading candidate. The main issue is catalysis of the methanol oxidation reaction, which is currently very sluggish and

leads to high overpotential at operating current densities. The side reactions produce carbon monoxide, which is a poison for the noble metal catalysts currently used. A second issue is the problem of methanol containment by the membrane at the negative electrode. The crossover of methanol leads to losses of fuel and reduced efficiency.

Many other battery and fuel-cell systems are under continuing development. Fuel cells operating at high temperatures have the advantage of improved electrode kinetics, but significant technical challenges include materials problems, especially corrosion and thermal management. Development of molten carbonate fuel cells (MCFCs) and solid oxide fuel cells (SOFCs) has been ongoing for several decades. The MCFC uses a eutectic mixture of lithium and potassium carbonates as the electrolyte. Because the MCFC operates at 650°C, reforming of a hydrocarbon fuel directly at the electrode is feasible. Steam reforming of methane followed by a shift-conversion reaction has been demonstrated in the MCFC. The SOFC operates in a range near 1000°C and is capable of internal reforming without a catalyst; however, because the reforming reaction is highly endothermic, thermal management is a problem. The yttria-stabilized zirconia electrolyte is an oxygen-anion-conducting electrolyte. Efforts are also under way to develop materials that are conductive at lower temperatures, at which materials problems are less severe.

In electrodeposition technology, damascene electroplating of copper was developed in the 1990s for chip interconnects, and copper is now displacing the aluminum-copper alloy for this purpose. This application of electroplating represents a major shift in the processing of on-chip wiring and has resulted in a 40% reduction in resistance of the interconnects. Damascene plating involves the deposition of a seed layer over a patterned insulating material. The electroplated material then covers the entire surface but fills trenches that serve as interconnects. Excess surface material is then removed through a planarization step such as chemical-mechanical polishing (CMP). Issues of metal distribution, plated copper voids, and copper diffusion into the insulator had to be overcome prior to commercial implementation.

Currently, adiponitrile is the only organic chemical produced in large quantity (10^8 kg/yr) by an electrochemical route. Other smaller-scale products include gluconic acid, piperidine, and *p*-aminophenol. Electroorganic syntheses in supercritical organic electrolytes have been demonstrated in bench-scale reactors. Production of dimethyl carbonate from the mixture-critical region was performed. There are at least a dozen electroorganic processes that are

reported in the pilot plant stage. The main attractions of electroorganic syntheses are high material and energy efficiency, ease of control, and ability to effect difficult oxidations or reductions. A general disadvantage is that a reaction at the counterelectrode must be performed, and unless a useful synthesis occurs there, cost benefits may not be realized.

Electrode materials must be capable of conducting electrons in the external circuit; therefore, metal and graphite are natural choices for electrode materials. Recently, doped polyacetylene has been used as an electrode material. Organic electrodes may be effective in reducing battery weight and significantly increasing specific energy. Semiconductor electrodes have been considered for use in the solar photolysis of water. Materials such as *n*-TiO₂ anodes and *p*-GaP cathodes have been successfully used to split water, but the efficiencies have been low.

The intrinsic performance advantages of hydrocarbon-based energy conversion systems are formidable. Currently, electrochemically based energy converters have found application in limited, niche markets. Electrochemical processes are frequently advantageous in terms of intrinsic efficiency, process control, and pollution reduction. Many systems await advances in electrocatalysis and materials.

SEE ALSO THE FOLLOWING ARTICLES

ALUMINUM • BATTERIES • CHEMICAL THERMODYNAMICS • ELECTROCHEMISTRY • FUEL CELLS, APPLICATIONS IN STATIONARY POWER SYSTEMS • KINETICS (CHEMISTRY) • TRANSPORTATION APPLICATIONS FOR FUEL CELLS

BIBLIOGRAPHY

- Appleby, A. J., and Foulkes, F. R., eds. (1993). "Fuel Cell Handbook," Krieger, Malabar, Fla.
- Newman, J. (1991). "Electrochemical Systems," 2nd ed, Prentice Hall, Englewood Cliffs, N.J.
- Pletcher, D., and Walsh, F. C. (1990). "Industrial Electrochemistry," 2nd ed., Chapman & Hall, London.
- Prentice, G. A. (1991). "Electrochemical Engineering Principles," Prentice Hall, Englewood Cliffs, N.J.
- Tobias, C. W., Delahay, P., and Gerischer, H., eds. (1961). "Advances in Electrochemistry and Electrochemical Engineering," Wiley (Interscience), New York.
- Varma, R., and Selman, J. R., eds. (1990). "Techniques for Characterization of Electrodes and Electrochemical Processes," Wiley, New York.



Fluid Dynamics (Chemical Engineering)

Richard W. Hanks

Brigham Young University

- I. Introduction
- II. Basic Field Equations (Differential or Microscopic)
- III. Basic Field Equations (Averaged or Macroscopic)
- IV. Laminar Flow
- V. Turbulent Flow
- VI. Applications

GLOSSARY

Field Mathematical representation of a physical quantity; at every point of space the mathematical quantity is defined as continuous for all necessary orders of differentiation.

Ground profile Plot of physical ground elevations along a pipeline route.

Head Any hydraulic energy quantity converted to an equivalent hydrostatic pressure and expressed as a column height of fluid.

Hydraulic grade line Graphic representation of the mechanical energy equation as hydraulic or pressure head against length; slope is frictional head loss per unit length.

Mixing length Mean distance over which a turbulent eddy retains its identity; phenomenological measure

of a turbulence length scale in a zero-parameter model.

Physical component Tensor component that has the physical dimensions of the property being represented.

Reynolds stress Nondiagonal element of the correlation dyad for fluctuation velocity components in a turbulent flow; commonly interpreted as a shear component of the extra stress caused by the turbulence.

Tensor Matrix operator that transforms one vector function into another; all tensorial functions and entities must transform properly according to laws of coordinate transformation and retain both formal and operational invariance.

THE MECHANICS OF FLUIDS is a broad subject dealing with all of the phenomena of fluid behavior. Subtended

within this subject is the subset of phenomena associated specifically with the kinematic and dynamic behavior of fluids. Kinematics is the study of motion per se, while dynamics includes the response of specific materials to applied forces. This requires one to apply the theory of deformable continuum fields. In its most general form the continuum field theory includes both fluid mechanics and dynamics in all their myriad forms. This article deals specifically with kinematic and dynamic applications.

I. INTRODUCTION

The phenomena of fluid mechanics are myriad and multi-form. In the practice of chemical engineering, most applications of fluid mechanics are associated with either flow through a bounded duct or flow around a fixed object in the context of design of processing equipment. The details of such problems may be very simple or extremely complex. The chemical engineer must know how to apply standard theoretical and empirical procedures to solve these problems. In cases where standard methods fail, he or she must also know how to apply fundamental principles and develop an appropriate solution. To this end this article deals with both the fundamentals and the application thereof to bounded duct flows and flows about objects of incompressible liquids of the type commonly encountered by practicing chemical engineers. The phenomena associated with compressible flow, two-phase gas–liquid flow, and flow through porous media are not considered because of space limitations.

II. BASIC FIELD EQUATIONS (DIFFERENTIAL OR MICROSCOPIC)

A. Generic Principle of Balance

The fundamental theory of fluid mechanics is expressed in the mathematical language of continuum tensor field calculus. An exhaustive treatment of this subject is found in the treatise by Truesdell and Toupin (1960). Two fundamental classes of equations are required: (1) the generic equations of balance and (2) the constitutive relations.

The generic equations of balance are statements of truth, which is *a priori* self-evident and which must apply to all continuum materials regardless of their individual characteristics. Constitutive relations relate diffusive flux vectors to concentration gradients through phenomenological parameters called transport coefficients. They describe the detailed response characteristics of specific materials. There are seven generic principles: (1) conservation of mass, (2) balance of linear momentum, (3) balance of ro-

tational momentum, (4) balance of energy, (5) conservation of charge–current, (6) conservation of magnetic flux, and (7) thermodynamic irreversibility.

In the vast majority of situations of importance to chemical engineers, the conservation of charge–current and magnetic flux are of no importance, and therefore, we will not consider them further here. They would be of considerable importance in a magnetohydrodynamic problem.

The four balance or conservation principles can all be represented in terms of a general equation of balance written in integral form as

$$\underbrace{\iiint_V \frac{\partial \psi}{\partial t} dV}_{\text{Net increase of } \psi \text{ in } V} = - \underbrace{\iint_S \psi \mathbf{v} \cdot \mathbf{n} ds}_{\text{Net convective influx of } \psi} - \underbrace{\iint_S \mathbf{j}_{D\psi} \cdot \mathbf{n} ds}_{\text{Net diffusive influx of } \psi} + \underbrace{\iiint_V \dot{r}_\psi dV}_{\text{Net production of } \psi \text{ in } V} \quad (1)$$

or in differential form as ($\mathbf{n}$ is the outward-directed normal vector; hence, $-\psi \mathbf{v} \cdot \mathbf{n} ds$ represents influx)

$$\underbrace{\frac{\partial \psi}{\partial t}}_{\text{Net increase of } \psi \text{ at point}} = - \underbrace{\nabla \cdot \psi \mathbf{v}}_{\text{Net convective influx of } \psi} - \underbrace{\nabla \cdot \mathbf{j}_{D\psi}}_{\text{Net diffusive influx of } \psi} + \underbrace{\dot{r}_\psi}_{\text{Net production of } \psi \text{ at point}} \quad (2)$$

where ψ represents the concentration or density of any transportable property of any tensorial order, $\mathbf{j}_{D\psi}$ represents the diffusive transport flux of property ψ , and $\dot{r}_\psi$ represents the volumetric rate of production or generation of property ψ within the volume V , which is bounded by the surface S .

Equation (2) is expressed in the Eulerian frame of reference, in which the volume element under consideration is fixed in space, and material is allowed to flow in and out of the element. An equivalent representation of very different appearance is the Lagrangian frame of reference, in which the volume element under consideration moves with the fluid and encapsulates a fixed mass of material so that no flow of mass in or out is permitted. In this frame of reference, Eq. (2) becomes

$$D\psi/Dt = -\psi \nabla \cdot \mathbf{v} - \nabla \cdot \mathbf{j}_{D\psi} + \dot{r}_\psi, \quad (3)$$

where the new differential term $D\psi/Dt$ is called the substantial or material derivative of ψ and is defined by the relation

$$\frac{D\psi}{Dt} = \frac{\partial \psi}{\partial t} + \mathbf{v} \cdot \nabla \psi. \quad (4)$$

Equations (2) and (3) are related by an obvious vector identity.

B. Equation of Continuity

If the generic property ψ is identified as the mass density ρ of a material, then Eq. (2) represents the generic principle of conservation of mass. The diffusive flux vector $\mathbf{j}_{D\rho}$ is equal to 0 and also $\dot{r}_\rho$ equals 0. Thus, the statement of conservation of mass, or equation of continuity, is

$$\partial\rho/\partial t = -\nabla \cdot \rho\mathbf{v} \quad (5)$$

in the Eulerian frame or

$$D\rho/Dt = -\rho\nabla \cdot \mathbf{v} \quad (6)$$

in the Lagrangian frame. The following are specific expressions for Eq. (5) in the three most commonly used systems:

Cartesian

$$-\frac{\partial\rho}{\partial t} = \frac{\partial}{\partial x}(\rho v_x) + \frac{\partial}{\partial y}(\rho v_y) + \frac{\partial}{\partial z}(\rho v_z) \quad (7)$$

Cylindrical Polar

$$-\frac{\partial\rho}{\partial t} = \frac{1}{r}\frac{\partial}{\partial r}(r\rho v_r) + \frac{1}{r}\frac{\partial}{\partial\theta}(\rho v_\theta) + \frac{\partial}{\partial z}(\rho v_z) \quad (8)$$

Spherical Polar

$$-\frac{\partial\rho}{\partial t} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2\rho v_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}[(\sin\theta)\rho v_\theta] + \frac{1}{r\sin\theta}\frac{\partial}{\partial\phi}(\rho v_\phi) \quad (9)$$

C. Equations of Motion

The vector quantity $\rho\mathbf{v}$ represents both the convective mass flux and the concentration of linear momentum. Its vector product $\mathbf{x} \times \rho\mathbf{v}$ with a position vector $\mathbf{x}$ from some axis of rotation represents the concentration of angular momentum about that axis. If $\mathbf{g} = -\nabla\Phi$ is an external body or action-at-a-distance force per unit mass, where Φ is a potential energy field, then the vector $\rho\mathbf{g}$ represents the volumetric rate of generation or production of linear momentum. The vector $\mathbf{x} \times \rho\mathbf{g}$ is the volumetric production rate of angular momentum.

Surface tractions or contact forces produce a stress field in the fluid element characterized by a stress tensor $\mathbf{T}$. Its negative is interpreted as the diffusive flux of momentum, and $\mathbf{x} \times (-\mathbf{T})$ is the diffusive flux of angular momentum or torque distribution. If stresses and torques are presumed to be in local equilibrium, the tensor $\mathbf{T}$ is easily shown to be symmetric.

When all of these quantities are introduced into Eq. (2), one obtains

$$\frac{\partial}{\partial t}(\rho\mathbf{v}) = -\nabla \cdot \rho\mathbf{v}\mathbf{v} - \nabla \cdot (-\mathbf{T}) + \rho\mathbf{g}, \quad (10)$$

which is known variously as Cauchy's equations of motion, Cauchy's first law of motion, the stress equations of motion, or Newton's second law for continuum fluids. Regardless of the name applied to Eq. (10), Truesdell and Toupin (1960) identify it and the statement of symmetry of $\mathbf{T}$ as the *fundamental equations of continuum mechanics*.

By using the vector identities relating Eulerian and Lagrangian frames together with the equation of continuity, one can convert Eq. (10) to an equivalent form:

$$\rho \frac{D\mathbf{v}}{Dt} = \rho\mathbf{g} - \nabla p - \nabla \cdot \boldsymbol{\tau}. \quad (11)$$

In this equation the stress tensor $\mathbf{T}$ has been partitioned into two parts in accordance with

$$\mathbf{T} = -p\boldsymbol{\delta} + \mathbf{P} = -p\boldsymbol{\delta} - \boldsymbol{\tau}, \quad (12)$$

where $-p$ is the mean normal stress defined by

$$-p = \frac{1}{3}(T_{xx} + T_{yy} + T_{zz}) \quad (13)$$

and $\mathbf{P}$ is known variously as the viscous stress tensor, the extra stress tensor, the shear stress tensor, or the stress deviator tensor. It contains both shear stresses (the off-diagonal elements) and normal stresses (the diagonal elements), both of which are related functionally to velocity gradient components by means of constitutive relations. In purely viscous fluids only the shear stresses are important, but the normal stresses become important when elasticity becomes a characteristic of the fluid. In incompressible liquids the mean normal stress is a dynamic parameter that replaces the thermodynamic pressure. It is the gradient of this pressure that is always dealt with in engineering design problems.

If one performs the vector operation $\mathbf{x} \times$ (equations of motion), the balance of rotational momentum or moment of momentum about an axis of rotation is obtained. It is this equation that forms the basis of design of rotating machinery such as centrifugal pumps and turbomachinery.

Equation (11) is written in the form of Newton's second law and states that the mass times acceleration of a fluid particle is equal to the sum of the forces causing that acceleration. In flow problems that are accelerationless ($D\mathbf{v}/Dt = 0$) it is sometimes possible to solve Eq. (11) for the stress distribution independently of any knowledge of the velocity field in the system. One special case where this useful feature of these equations occurs is the case of rectilinear pipe flow. In this special case the solution of complex fluid flow problems is greatly simplified because the stress distribution can be discovered before the constitutive relation must be introduced. This means that only a first-order differential equation must be solved rather than a second-order (and often nonlinear) one. The following are the components of Eq. (11) in rectangular Cartesian, cylindrical polar, and spherical polar coordinates:

Cartesian:

x Component

$$\rho \left(\frac{\partial v_x}{\partial t} + v_x \frac{\partial v_x}{\partial x} + v_y \frac{\partial v_x}{\partial y} + v_z \frac{\partial v_x}{\partial z} \right) = -\frac{\partial p}{\partial x} - \left(\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{yx}}{\partial y} + \frac{\partial \tau_{zx}}{\partial z} \right) + \rho g_x \quad (14)$$

y Component

$$\rho \left(\frac{\partial v_y}{\partial t} + v_x \frac{\partial v_y}{\partial x} + v_y \frac{\partial v_y}{\partial y} + v_z \frac{\partial v_y}{\partial z} \right) = -\frac{\partial p}{\partial y} - \left(\frac{\partial \tau_{xy}}{\partial x} + \frac{\partial \tau_{yy}}{\partial y} + \frac{\partial \tau_{zy}}{\partial z} \right) + \rho g_y \quad (15)$$

z Component

$$\rho \left(\frac{\partial v_z}{\partial t} + v_x \frac{\partial v_z}{\partial x} + v_y \frac{\partial v_z}{\partial y} + v_z \frac{\partial v_z}{\partial z} \right) = -\frac{\partial p}{\partial z} - \left(\frac{\partial \tau_{xz}}{\partial x} + \frac{\partial \tau_{yz}}{\partial y} + \frac{\partial \tau_{zz}}{\partial z} \right) + \rho g_z \quad (16)$$

Cylindrical Polar:

r Component

$$\rho \left(\frac{\partial v_r}{\partial t} + v_r \frac{\partial v_r}{\partial r} + \frac{v_\theta}{r} \frac{\partial v_r}{\partial \theta} - \frac{v_\theta^2}{r} + v_z \frac{\partial v_r}{\partial z} \right) = -\frac{\partial p}{\partial r} - \left(\frac{1}{r} \frac{\partial}{\partial r} (r \tau_{rr}) + \frac{1}{r} \frac{\partial \tau_{r\theta}}{\partial \theta} - \frac{\tau_{\theta\theta}}{r} + \frac{\partial \tau_{rz}}{\partial z} \right) + \rho g_r \quad (17)$$

θ Component

$$\rho \left(\frac{\partial v_\theta}{\partial t} + v_r \frac{\partial v_\theta}{\partial r} + \frac{v_\theta}{r} \frac{\partial v_\theta}{\partial \theta} + \frac{v_r v_\theta}{r} + v_z \frac{\partial v_\theta}{\partial z} \right) = -\frac{1}{r} \frac{\partial p}{\partial \theta} - \left(\frac{1}{r^2} \frac{\partial}{\partial r} (r^2 \tau_{r\theta}) + \frac{1}{r} \frac{\partial \tau_{\theta\theta}}{\partial \theta} + \frac{\partial \tau_{\theta z}}{\partial z} \right) + \rho g_\theta \quad (18)$$

z Component

$$\rho \left(\frac{\partial v_z}{\partial t} + v_r \frac{\partial v_z}{\partial r} + \frac{v_\theta}{r} \frac{\partial v_z}{\partial \theta} + v_z \frac{\partial v_z}{\partial z} \right) = -\frac{\partial p}{\partial z} - \left(\frac{1}{r} \frac{\partial}{\partial r} (r \tau_{rz}) + \frac{1}{r} \frac{\partial \tau_{\theta z}}{\partial \theta} + \frac{\partial \tau_{zz}}{\partial z} \right) + \rho g_z \quad (19)$$

Spherical Polar:

r Component

$$\rho \left(\frac{\partial v_r}{\partial t} + v_r \frac{\partial v_r}{\partial r} + \frac{v_\theta}{r} \frac{\partial v_r}{\partial \theta} + \frac{v_\phi}{r \sin \theta} \frac{\partial v_r}{\partial \phi} - \frac{v_\theta^2 + v_\phi^2}{r} \right) = -\frac{\partial p}{\partial r} - \left(\frac{1}{r^2} \frac{\partial}{\partial r} (r^2 \tau_{rr}) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\tau_{r\theta} \sin \theta) + \frac{1}{r \sin \theta} \frac{\partial \tau_{r\phi}}{\partial \phi} - \frac{\tau_{\theta\theta} + \tau_{\phi\phi}}{r} \right) + \rho g_r \quad (20)$$

θ Component

$$\rho \left(\frac{\partial v_\theta}{\partial t} + v_r \frac{\partial v_\theta}{\partial r} + \frac{v_\theta}{r} \frac{\partial v_\theta}{\partial \theta} + \frac{v_\phi}{r \sin \theta} \frac{\partial v_\theta}{\partial \phi} + \frac{v_r v_\theta}{r} - \frac{v_\phi^2 \cot \theta}{r} \right) = -\frac{1}{r} \frac{\partial p}{\partial \theta} - \left(\frac{1}{r^2} \frac{\partial}{\partial r} (r^2 \tau_{r\theta}) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (\tau_{\theta\theta} \sin \theta) + \frac{1}{r \sin \theta} \frac{\partial \tau_{\theta\phi}}{\partial \phi} + \frac{\tau_{r\theta}}{r} - \frac{\cot \theta}{r} \tau_{\phi\phi} \right) + \rho g_\theta \quad (21)$$

ϕ Component

$$\rho \left(\frac{\partial v_\phi}{\partial t} + v_r \frac{\partial v_\phi}{\partial r} + \frac{v_\theta}{r} \frac{\partial v_\phi}{\partial \theta} + \frac{v_\phi}{r \sin \theta} \frac{\partial v_\phi}{\partial \phi} + \frac{v_\phi v_r}{r} + \frac{v_\theta v_\phi}{r} \cot \theta \right) = -\frac{1}{r \sin \theta} \frac{\partial p}{\partial \phi} - \left(\frac{1}{r^2} \frac{\partial}{\partial r} (r^2 \tau_{r\phi}) + \frac{1}{r} \frac{\partial \tau_{\theta\phi}}{\partial \theta} \right) + \frac{1}{r \sin \theta} \frac{\partial \tau_{\phi\phi}}{\partial \phi} + \frac{\tau_{r\phi}}{r} + \frac{2 \cot \theta}{r} \tau_{\theta\phi} \right) + \rho g_\phi \quad (22)$$

Two terms in Eqs. (17) and (18) are worthy of special note. In Eq. (17) the term $\rho v_\theta^2/r$ is the centrifugal “force.” That is, it is the effective force in the r direction arising from fluid motion in the θ direction. Similarly, in Eq. (18) $\rho v_r v_\theta/r$ is the Coriolis force, or effective force in the θ direction due to motion in both the r and θ directions. Both of these forces arise naturally in the transformation of coordinates from the Cartesian frame to the cylindrical polar frame. They are properly part of the acceleration vector and do not need to be added on physical grounds.

D. Total Energy Balance

Two types of energy terms must be considered: (1) thermal and (2) mechanical. The specific internal energy is $u = C_v T$, where C_v is the heat capacity and T is the temperature of the fluid. The specific kinetic energy is $v^2/2$. Thus, the total energy density is $\rho(u + v^2/2)$. Thermal energy diffuses into the fluid by means of a heat flux vector $\mathbf{q}$. Mechanical energy diffuses in by means of work done against the stresses $\mathbf{v} \cdot (-\mathbf{T})$. Energy may be produced internally in the fluid by chemical reactions at a rate $\dot{r}_{CR}$ and by the action of external body forces $\mathbf{v} \cdot \rho \mathbf{g}$. Thus, Eq. (2) can be written as Eulerian-form total energy balance as

$$\begin{aligned} \frac{\partial}{\partial t} \left[\rho \left(u + \frac{v^2}{2} \right) \right] = & -\nabla \cdot \left[\rho \left(u + \frac{v^2}{2} \right) \mathbf{v} \right] \\ & - \nabla \cdot \mathbf{q} - \nabla \cdot [\mathbf{v} \cdot (-\mathbf{T})] \\ & + \mathbf{v} \cdot \rho \mathbf{g} + \dot{r}_{CR}. \end{aligned} \quad (23)$$

By appropriate manipulation as before, this can be written in Lagrangian form as

$$\begin{aligned} \rho \frac{D}{Dt} \left(u + \frac{v^2}{2} \right) = & -\nabla \cdot \mathbf{q} - \nabla \cdot [\mathbf{v} \cdot (-\mathbf{T})] \\ & + \mathbf{v} \cdot \rho \mathbf{g} + \dot{r}_{CR}. \end{aligned} \quad (24)$$

By using Eq. (12) the term $-\nabla \cdot [\mathbf{v} \cdot (-\mathbf{T})]$ can be written as

$$-\nabla \cdot [\mathbf{v} \cdot (-\mathbf{T})] = -\nabla \cdot (p\mathbf{v}) - \mathbf{v} \cdot (\nabla \cdot \boldsymbol{\tau}) - \boldsymbol{\tau} : \nabla \mathbf{v}. \quad (25)$$

In Eq. (25) the term $\mathbf{v} \cdot (\nabla \cdot \boldsymbol{\tau})$ represents reversible stress work, while $\boldsymbol{\tau} : \nabla \mathbf{v}$ represents irreversible or entropy-producing stress work. The following are expressions for the latter quantity in rectangular Cartesian, cylindrical polar, and spherical polar coordinates:

Cartesian

$$\begin{aligned} (\boldsymbol{\tau} : \nabla \mathbf{v}) = & \tau_{xx} \left(\frac{\partial v_x}{\partial x} \right) + \tau_{yy} \left(\frac{\partial v_y}{\partial y} \right) + \tau_{zz} \left(\frac{\partial v_z}{\partial z} \right) \\ & + \tau_{xy} \left(\frac{\partial v_x}{\partial y} + \frac{\partial v_y}{\partial x} \right) + \tau_{yz} \left(\frac{\partial v_y}{\partial z} + \frac{\partial v_z}{\partial y} \right) \\ & + \tau_{zx} \left(\frac{\partial v_z}{\partial x} + \frac{\partial v_x}{\partial z} \right) \end{aligned} \quad (26)$$

Cylindrical Polar

$$\begin{aligned} (\boldsymbol{\tau} : \nabla \mathbf{v}) = & \tau_{rr} \left(\frac{\partial v_r}{\partial r} \right) + \tau_{\theta\theta} \left(\frac{1}{r} \frac{\partial v_\theta}{\partial \theta} + \frac{v_r}{r} \right) + \tau_{zz} \left(\frac{\partial v_z}{\partial z} \right) \\ & + \left[\tau_{r\theta} r \frac{\partial}{\partial r} \left(\frac{v_\theta}{r} \right) + \frac{1}{r} \frac{\partial v_r}{\partial \theta} \right] \\ & + \tau_{\theta z} \left(\frac{1}{r} \frac{\partial v_z}{\partial \theta} + \frac{\partial v_\theta}{\partial z} \right) + \tau_{rz} \left(\frac{\partial v_z}{\partial r} + \frac{\partial v_r}{\partial z} \right) \end{aligned} \quad (27)$$

Spherical Polar

$$\begin{aligned} (\boldsymbol{\tau} : \nabla \mathbf{v}) = & \tau_{rr} \left(\frac{\partial v_r}{\partial r} \right) + \tau_{\theta\theta} \left(\frac{1}{r} \frac{\partial v_\theta}{\partial \theta} + \frac{v_r}{r} \right) \\ & + \tau_{\phi\phi} \left(\frac{1}{r \sin \theta} \frac{\partial v_\phi}{\partial \phi} + \frac{v_r}{r} + \frac{v_\theta \cot \theta}{r} \right) \\ & + \tau_{r\theta} \left(\frac{\partial v_\theta}{\partial r} + \frac{1}{r} \frac{\partial v_r}{\partial \theta} - \frac{v_\theta}{r} \right) \\ & + \tau_{r\phi} \left(\frac{\partial v_\phi}{\partial r} + \frac{1}{r \sin \theta} \frac{\partial v_r}{\partial \phi} - \frac{v_\phi}{r} \right) \\ & + \tau_{\theta\phi} \frac{1}{r} \left(\frac{\partial v_\phi}{\partial \theta} + \frac{1}{r \sin \theta} \frac{\partial v_\theta}{\partial \phi} - \frac{\cot \theta}{r} v_\phi \right) \end{aligned} \quad (28)$$

Equation (23) represents the total energy balance or first law of thermodynamics. It includes all forms of energy transport. An independent energy equation, which does not represent a generic balance relation, is obtained by performing the operation $\mathbf{v} \cdot$ (equations of motion) and is

$$\rho \frac{D}{Dt} \frac{v^2}{2} = -\mathbf{v} \cdot \nabla p - \mathbf{v} \cdot (\nabla \cdot \boldsymbol{\tau}) + \mathbf{v} \cdot \rho \mathbf{g}. \quad (29)$$

This relation, called the mechanical energy equation, describes the rate of increase of kinetic energy in a fluid element as a result of the action of external body forces, pressure, and reversible stress work.

When Eq. (29) is subtracted from Eq. (24), one obtains

$$\rho \frac{Du}{Dt} = -\nabla \cdot \mathbf{q} - p \nabla \cdot \mathbf{v} = \boldsymbol{\tau} : \nabla \mathbf{v} + \dot{r}_{CR}, \quad (30)$$

which is called the thermal energy equation. It describes the rate of increase of thermal internal energy of a fluid element by the action of heat fluxes, chemical reactions, volumetric expansion of the fluid, and irreversible stress work.

Clearly, only two of the three energy equations are independent, the third being obtained by sum or difference from the first two. The coupling between Eqs. (29) and (30) occurs by means of Eq. (25), which represents the total work done on the fluid element by the stress field. Neither Eq. (29) nor Eq. (30) is a balance relation by itself, but the sum of the two, Eq. (24), is.

E. Entropy Production Principle

We cannot write down *a priori* a generic balance relation for the entropy of a fluid. We can, however, derive a result that can be placed in the same form as Eq. (3) and therefore recognized as a balance relation. By working with the combined first and second laws of thermodynamics, one

can show that the rate of increase of specific entropy is given by

$$\rho \frac{Ds}{Dt} = -\nabla \cdot \frac{\mathbf{q}}{T} + \frac{1}{T} \left(-\boldsymbol{\tau} : \nabla \mathbf{v} + \frac{1}{T} \mathbf{q} \cdot \nabla T + \dot{r}_{CR} \right), \quad (31)$$

where s is the specific entropy. Equation (31) is in the Lagrangian form of Eq. (3) with $\psi = \rho s$ and where the equation of continuity has been invoked. Thus, we recognize the term $-\nabla \cdot (\mathbf{q}/T)$ as the diffusive influx of entropy and the production or generation of entropy as the remaining three terms on the right side of Eq. (31). In the absence of chemical reactions, the principle of entropy production (or “postulate of irreversibility,” as Truesdell has called it) states that

$$\frac{1}{T} (-\boldsymbol{\tau} : \nabla \mathbf{v}) + \frac{1}{T^2} \mathbf{q} \cdot \nabla T \geq 0. \quad (32)$$

From Eq. (32) it follows that only the part $-\boldsymbol{\tau} : \nabla \mathbf{v}$ of the stress work contributes to the production of entropy; hence, it is the “irreversible” or nonrecoverable work. The remainder of the stress work, expressed by $\mathbf{v} \cdot (\nabla \cdot \mathbf{T})$, is “reversible” or recoverable, as already described.

F. Constitutive Relations

The generic balance relations and the derived relations presented in the preceding section contain various diffusion flux tensors. Although the equation of continuity as presented does not contain a diffusion flux vector, were it to have been written for a multicomponent mixture, there would have been such a diffusion flux vector. Before any of these equations can be solved for the various field quantities, the diffusion fluxes must be related to gradients in the field potentials ϕ .

In general, the fluxes are related to gradients of the specific concentrations by relations of the form

$$\mathbf{j}_{D\psi} = -\beta \nabla \phi \quad (33)$$

or

$$\mathbf{j}_{d\psi} = -\mathbf{B} \cdot \nabla \phi. \quad (34)$$

In the form of Eq. (33) β is a scalar parameter called a transport coefficient. In the form of Eq. (34) $\mathbf{B}$ is a tensor, the elements of which are the transport coefficients. In either form the transport coefficients may be complex nonlinear functions of the scalar invariants of $\nabla \phi$.

For isotropic fluids the heat flux vector $\mathbf{q}$ takes the form

$$\mathbf{q} = -k_T \nabla T, \quad (35)$$

where k_T is the thermal conductivity. Equation (35) is known as Fourier’s law of conduction. The momentum flux tensor $\boldsymbol{\tau}$ is expressed in the form

$$\boldsymbol{\tau} = -2\mu_a \mathbf{D}, \quad (36)$$

where μ_a is the apparent viscosity or viscosity function and $\mathbf{D}$ is the symmetric part of $\nabla \mathbf{v}$ given by

$$\mathbf{D} = \frac{1}{2}(\nabla \mathbf{v} + \nabla \mathbf{v}^T). \quad (37)$$

In general, μ_a is a complex and often nonlinear function of Π_D , the second principal invariant of $\mathbf{D}$; Π_D is given by

$$-\Pi_D = \frac{1}{2}[(\nabla \cdot \mathbf{v})^2 - \mathbf{D} : \mathbf{D}]. \quad (38)$$

In the special case of a Newtonian fluid, $\mu_a = \mu$ is a constant called the viscosity of the fluid and Eq. (36) becomes Newton’s “law” of viscosity. In a great many practical cases of interest to chemical engineers, however, the non-Newtonian form of Eq. (36) is encountered.

The formulation of proper constitutive relations is a complex problem and is the basis of the science of rheology, which cannot be covered here. This section presents only four relatively simple constitutive relations that have proved to be practically useful to chemical engineers. Elastic fluid behavior is expressly excluded from consideration. The following equations are a listing of these constitutive relations; many others are possible:

Bingham Plastics

$$\boldsymbol{\tau} = -2 \left\{ \mu_\infty \pm \frac{\tau_0}{2\sqrt{-2\Pi_D}} \right\} \mathbf{D}, \quad \frac{1}{2} \boldsymbol{\tau} : \boldsymbol{\tau} > \tau_0^2 \quad (39)$$

$$0 = \mathbf{D}, \quad \frac{1}{2} \boldsymbol{\tau} : \boldsymbol{\tau} \leq \tau_0^2 \quad (40)$$

Ostwald–DeWael or Power Law

$$\boldsymbol{\tau} = -2k |2\sqrt{-2\Pi_D}|^{n-1} \mathbf{D} \quad (41)$$

Herschel–Bulkley or Yield Power Law

$$\boldsymbol{\tau} = -2 \left\{ k |2\sqrt{-2\Pi_D}|^{n-1} \pm \frac{\tau_0}{2\sqrt{-2\Pi_D}} \right\} \mathbf{D} \quad \frac{1}{2} \boldsymbol{\tau} : \boldsymbol{\tau} > \tau_0^2 \quad (42)$$

$$0 = \mathbf{D}, \quad \frac{1}{2} \boldsymbol{\tau} : \boldsymbol{\tau} \leq \tau_0^2 \quad (43)$$

Casson

$$\frac{\boldsymbol{\tau}}{|2 - 2\Pi_\tau|^{1/2}} = -2 \frac{\pm \tau_0}{|2 - 2\Pi_D|} + \frac{\mu_\infty}{|2 - 2\Pi_D|^{1/2}} \mathbf{D} \quad \frac{1}{2} \boldsymbol{\tau} : \boldsymbol{\tau} > \tau_0^2 \quad (44)$$

$$0 = \mathbf{D}, \quad \frac{1}{2} \boldsymbol{\tau} : \boldsymbol{\tau} \leq \tau_0^2 \quad (45)$$

When these constitutive relations are coupled with the stress distributions derived from the equations of motion, details of the velocity fields can be calculated, as can the overall relation between pressure drop and volume flow rate.

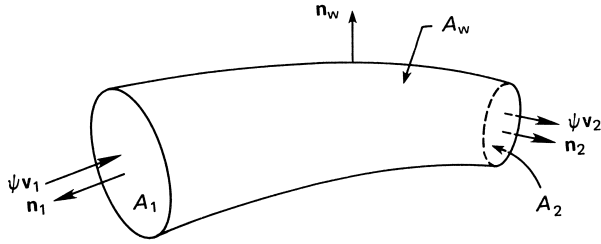


FIGURE 1 Schematic illustration of notation used in developing macroscopic equations.

III. BASIC FIELD EQUATIONS (AVERAGED OR MACROSCOPIC)

While the differential equations presented here are general and can be used to solve all types of fluid mechanics problems, to the average “practical” chemical engineer they are often unintelligible and intimidating. Much more familiar to most engineers are the averaged or macroscopic forms of these equations.

Equation (1) contains the integral form of the general balance relation. In this form it is a Eulerian result. If we take the volume in question to be the entire volume of the pipe located between two planes located at points 1 and 2 separated by some finite distance, as shown in Fig. 1, Eq. (1) can be written in the following average or macroscopic form,

$$\frac{\partial \Psi}{\partial t} = -\langle \psi \mathbf{v} \cdot \mathbf{n} \rangle_2 A_2 - \langle \psi \mathbf{v} \cdot \mathbf{n} \rangle_1 A_1 - \langle \mathbf{j}_{D\psi} \cdot \mathbf{n} \rangle_2 A_2 - \langle \mathbf{j}_{D\psi} \cdot \mathbf{n} \rangle_1 A_1 - \langle \mathbf{j}_{D\psi} \cdot \mathbf{n} \rangle_w A_w + \dot{R}_\psi V, \quad (46)$$

where Ψ is the total content of ψ in volume V and $\dot{R}_\psi$ is the volume average rate of production of ψ in V . In this relation the caret brackets have the significance

$$\langle (\cdot) \cdot \mathbf{n} \rangle_k \equiv \frac{1}{A_k} \iint_{A_k} [(\cdot) \cdot \mathbf{n}]_k ds, \quad (47)$$

which is simply a statement of the mean value theorem of calculus applied to the integral in question. Equation (46) is an averaged or macroscopic form of the general balance relation and can be applied to mass, momentum, and energy.

A. Equation of Continuity

As before, there are no generation or diffusion terms for mass, so Eq. (46) becomes

$$\frac{\partial m}{\partial t} = \rho(\langle v \rangle_1 A_1 - \langle v \rangle_2 A_2). \quad (48)$$

The vast majority of practical chemical engineering problems are in steady-state operation, so that Eq. (48) reduces

simply to the statement that mass flow or volume flow is constant,

$$\langle v \rangle_1 A_1 = \langle v \rangle_2 A_2 = Q, \quad (49)$$

where Q is the volume flow. This relation defines the area mean velocity as Q/A . Equation (49) is the working form most often used.

B. Momentum Balance

Setting ψ equal to $\rho \mathbf{v}$ in Eq. (46) produces the macroscopic momentum balance. The term $\langle \mathbf{j}_{D\psi} \cdot \mathbf{n} \rangle_w$ represents the reaction force of the wall of the pipe on the fluid arising from friction and changes in the direction of flow. The term $\dot{R}_\psi V$ represents the action of the body force $\rho \mathbf{g}$ on the total flow. Thus, Eq. (46) becomes

$$\frac{\partial \mathbf{M}}{\partial t} = -\rho \langle \mathbf{v} \mathbf{v} \cdot \mathbf{n} \rangle_1 A_1 - \rho \langle \mathbf{v} \mathbf{v} \cdot \mathbf{n} \rangle_2 A_2 - \langle p \mathbf{n} \rangle_1 A_1 - \langle p \mathbf{n} \rangle_2 A_2 + \mathbf{F}_w + \rho V \mathbf{g}, \quad (50)$$

where $\mathbf{M}$ is the total momentum of the flow. Equation (50) can be solved at steady state for the reaction force $\mathbf{F}_w$ as

$$\mathbf{F}_w = \langle p \mathbf{n} \rangle_1 A_1 + \langle p \mathbf{n} \rangle_2 A_2 + \rho \langle \mathbf{v} \mathbf{v} \cdot \mathbf{n} \rangle_1 A_1 + \rho \langle \mathbf{v} \mathbf{v} \cdot \mathbf{n} \rangle_2 A_2 - \rho V \mathbf{g}. \quad (51)$$

As an illustration of the use of this result, consider the pipe bend shown schematically in Fig. 2. Presuming the pipe to lie entirely in the x - y plane, we compute $F_{wx} = \mathbf{i} \cdot \mathbf{F}_w$, $F_{wy} = \mathbf{j} \cdot \mathbf{F}_w$, and $F_{wz} = \mathbf{k} \cdot \mathbf{F}_w$ as follows:

$$F_{wx} = -p_1 A_1 \cos \phi_1 + p_2 A_2 \cos \phi_2 - \rho \langle v \rangle_1^2 A_1 \cos \phi_1 + \rho \langle v \rangle_2^2 A_2 \cos \phi_2 \quad (52)$$

$$F_{wy} = -p_1 A_1 \sin \phi_1 + p_2 A_2 \sin \phi_2 - \rho \langle v \rangle_1^2 A_1 \sin \phi_1 + \rho \langle v \rangle_2^2 A_2 \sin \phi_2 \quad (53)$$

$$F_{wz} = \rho V g \quad (54)$$

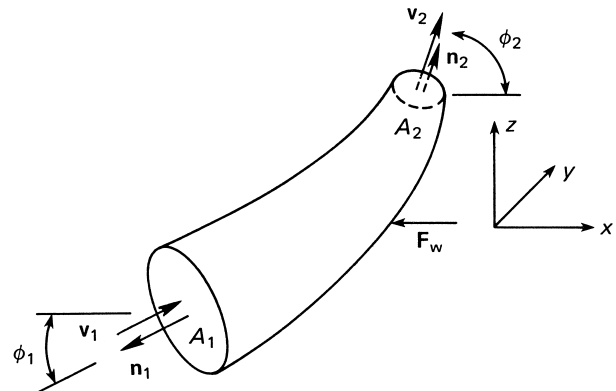


FIGURE 2 Illustration of forces on a pipe bend.

Thus, $(F_{wx}^2 + F_{wy}^2 + F_{wz}^2)^{1/2}$ is the magnitude of the force that would act in a bracing strut applied to the outside of the pipe bend to absorb the forces caused by turning the stream.

C. Energy Equations

When Eq. (46) is applied to energy quantities, a very large number of equivalent representations of the results are possible. Because of space limitations, we include only one commonly used variation here.

1. Total Energy (First Law of Thermodynamics)

When the various energy quantities used in arriving at Eq. (23) are introduced into Eq. (46), we obtain

$$\frac{\partial E}{\partial t} + \langle \rho e' \mathbf{v} \cdot \mathbf{n} \rangle_1 A_1 + \langle \rho e' \mathbf{v} \cdot \mathbf{n} \rangle_2 A_2 = \dot{Q} - \dot{W} + \dot{Q}'_{CR}, \quad (55)$$

in which $E = u + v^2/2 + \Phi$ is the total energy content of the fluid, $e' = e + p/\rho$, $\dot{Q}$ is the total thermal energy transfer rate,

$$\dot{Q} = - \iint_S \mathbf{q} \cdot \mathbf{n} ds, \quad (56)$$

$\dot{Q}'_{CR}$ is the total volumetric energy production rate due to chemical reactions or other such sources, and $\dot{W}$ is the total rate of work done or power expended against the viscous stresses,

$$\dot{W} = \iint_S (\mathbf{v} \cdot \mathbf{T}) \cdot \mathbf{n} ds. \quad (57)$$

In common engineering practice Eq. (55) is applied to steady flow in straight pipes and is divided by the mass flow rate $\dot{m} = \rho \langle v \rangle A$ to put it on a per unit mass basis,

$$\Delta u + \Delta \langle v \rangle^2/2 + g \Delta z + \Delta p/\rho = \hat{q} - \hat{w} + \hat{q}', \quad (58)$$

where the operator Δ implies average quantities at the downstream point minus the same average quantities at the upstream point. The terms on the right-hand side of Eq. (58) are just those on the right-hand side of Eq. (55) divided by $\rho \langle v \rangle A$. In Eq. (58) z is vertical elevation above an arbitrary datum plane.

2. Mechanical Energy (Bernoulli's Equation)

By considering Cauchy's equations of motion [Eq. (10)], Truesdell derived the theorem of stress means,

$$\begin{aligned} \iint_S G \mathbf{T} \cdot \mathbf{n} ds &= \iiint_V \mathbf{T} \cdot \nabla G dV + \iiint_V \rho G \frac{D\mathbf{v}}{Dt} dV \\ &- \iiint_V \rho G g dV, \end{aligned} \quad (59)$$

where G is a functional operator of any tensorial order and the other terms have the significance already described. In particular, if one sets G equal to $\mathbf{v} \cdot$, Eq. (59) results in

$$\frac{\partial}{\partial t} \frac{v^2}{2} + \langle \rho K' \mathbf{v} \cdot \mathbf{n} \rangle_1 A_1 + \langle \rho K' \mathbf{v} \cdot \mathbf{n} \rangle_2 A_2 = -\dot{W} - \dot{F}, \quad (60)$$

which is the macroscopic form of Eq. (29), the mechanical energy equation. In this expression $K' = e - u$ is the combined kinetic, potential, and pressure energy of the fluid; $\dot{F}$ is the energy dissipated by friction and is given by

$$\dot{F} = - \iiint_V \boldsymbol{\tau} : \nabla \mathbf{v} dV. \quad (61)$$

Consideration of Eqs. (29) and (32) shows that the mechanical energy equation involves only the recoverable or reversible work. In order to calculate this term on the average, however, it is necessary to compute the total work done $\dot{W}$ and subtract from it the part lost due to friction or the irreversible work $\dot{F}$. If Eq. (60) is applied to steady flow in a pipe and divided by the mass flow rate, the following per unit mass form is obtained,

$$\Delta \langle v \rangle^2/2 + g \Delta z + \Delta p/\rho = -\hat{w} - \hat{w}_f, \quad (62)$$

where $\hat{w}_f = \dot{F}/\rho \langle v \rangle A$ is the frictional energy loss per unit mass, and all other terms have the same significance as in Eq. (58). In practical engineering problems the key to the use of Eq. (62) is determining a numerical value for $\hat{w}_f$.

As we have seen, the above are variations of the mechanical energy equation. They are variously called the Bernoulli equation, the extended Bernoulli equation, or the engineering Bernoulli equation by writers of elementary fluid mechanics textbooks. Regardless of one's taste in nomenclature, Eq. (62) lies at the heart of nearly all practical engineering design problems.

a. Head concept. If Eq. (62) is divided by g , the gravitational acceleration constant, we obtain

$$\Delta \langle v \rangle^2/2g + \Delta z + \Delta p/\rho g = -h_s - h_f. \quad (63)$$

It will be observed that each term in Eq. (63) has physical dimensions of length. For example, if flow ceases, Eq. (63) reduces to

$$\Delta z + \Delta p/\rho g = 0, \quad (64)$$

which is just the equation of hydrostatic equilibrium and shows that the pressure differential existing between points 1 and 2 is simply the hydrostatic pressure due to a column of fluid of height $-\Delta z$. In a general situation each of the terms in Eq. (63) has the physical significance that it is the equivalent hydrostatic pressure "head" or height to which the respective type of energy term could be converted. Thus, $\Delta \langle v \rangle^2/2g$ is the velocity head, $\Delta p/\rho g$ is the

pressure head, $-h_s$ is the pump or shaft work head, h_f is the friction head, and Δz is the potential or ground head.

b. Friction head. In order to solve problems using Eq. (63), additional information is required regarding the nature of the friction head loss term $-h_f$. This information can be obtained by empirical correlation of experimental data, by theoretical solution of the field equations, or a combination of both. It is customary to express the friction head loss term as a proportionality with the dimensionless length of the pipe L/D and the velocity head in the pipe $\langle v \rangle^2/2g$,

$$h_f = f \frac{L}{D} \frac{\langle v \rangle^2}{2g}, \quad (65)$$

where f is called a friction factor. The problem is thus reduced to finding a functional relation between the dimensionless factor f and whatever variables with which it may be found to correlate. In practice, two definitions of the friction factor are in common use. The expression given in Eq. (65) is the Darcy–Weisbach form common to civil and mechanical engineering usage. An alternative form, commonly used by chemical engineers in the older literature, is the Fanning friction factor,

$$f' = f/4. \quad (66)$$

Care should always be exercised in using friction factors derived from a chart, table, or correlating equation to determine which type of friction factor is being obtained. The Darcy–Weisbach form is gradually supplanting the Fanning form as a consequence of most modern textbooks on fluid mechanics being written by either civil or mechanical engineers. Figure 3 is the widely accepted correlation for f for Newtonian fluids. This is called the Moody diagram. In it f is correlated as a function of the two dimensionless variables ε/D and $Re = D\langle v \rangle \rho / \mu$, where ε is a relative roughness factor expressed as an average depth of pit or height of protrusion on the wall of a rough pipe and Re is called the Reynolds number. Re is a dynamic similarity parameter. This means that two flows having the same value of Re are dynamically similar to one another. All variables in two pipes therefore scale in similar proportion to their Re values. The Moody diagram does not work for non-Newtonian fluids. In this case other methods, to be discussed below, must be employed.

c. Pump or work head. The pump head term in Eq. (63) is given by $h_s = \hat{w}_s/g$ and represents the head

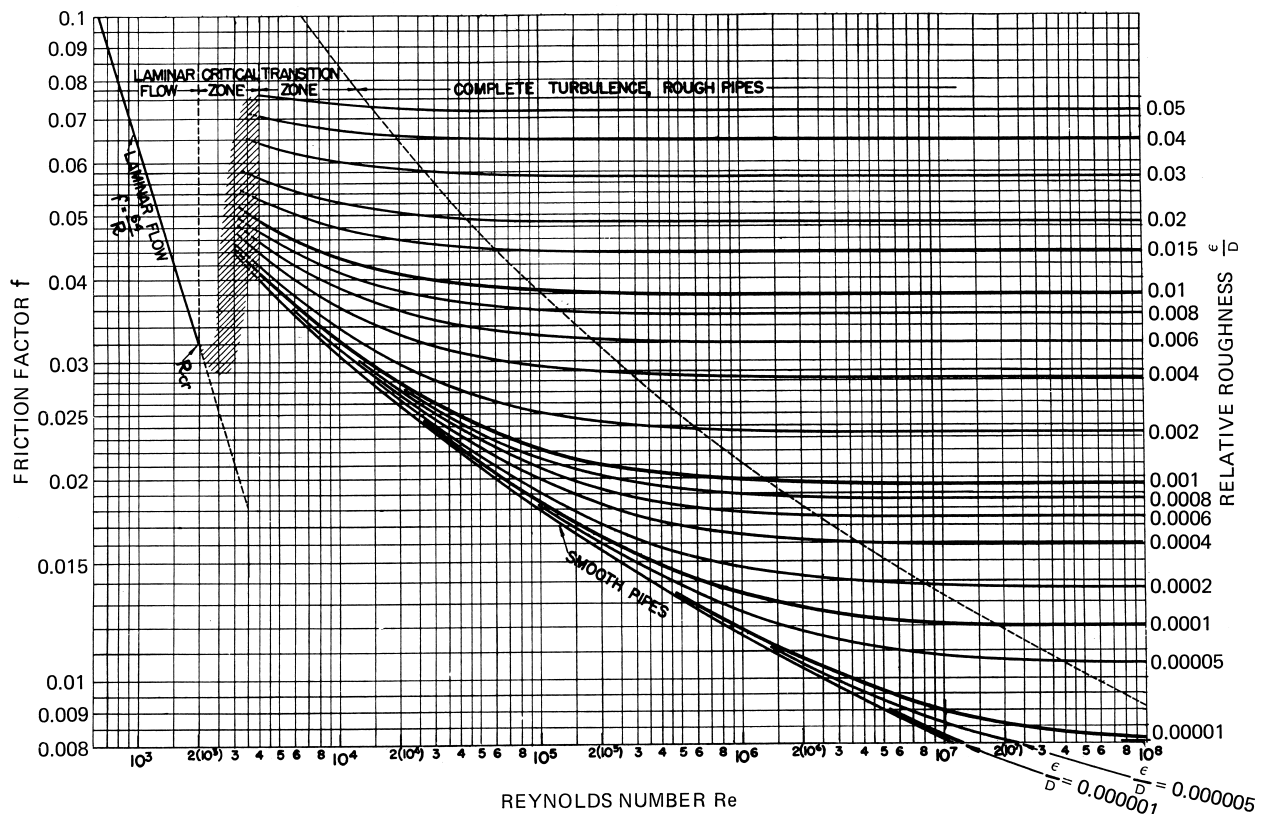


FIGURE 3A Moody diagram for Newtonian pipe flow. [Adapted from Moody, L. F. (1944). *Trans. ASME* **66**, 671–684.]

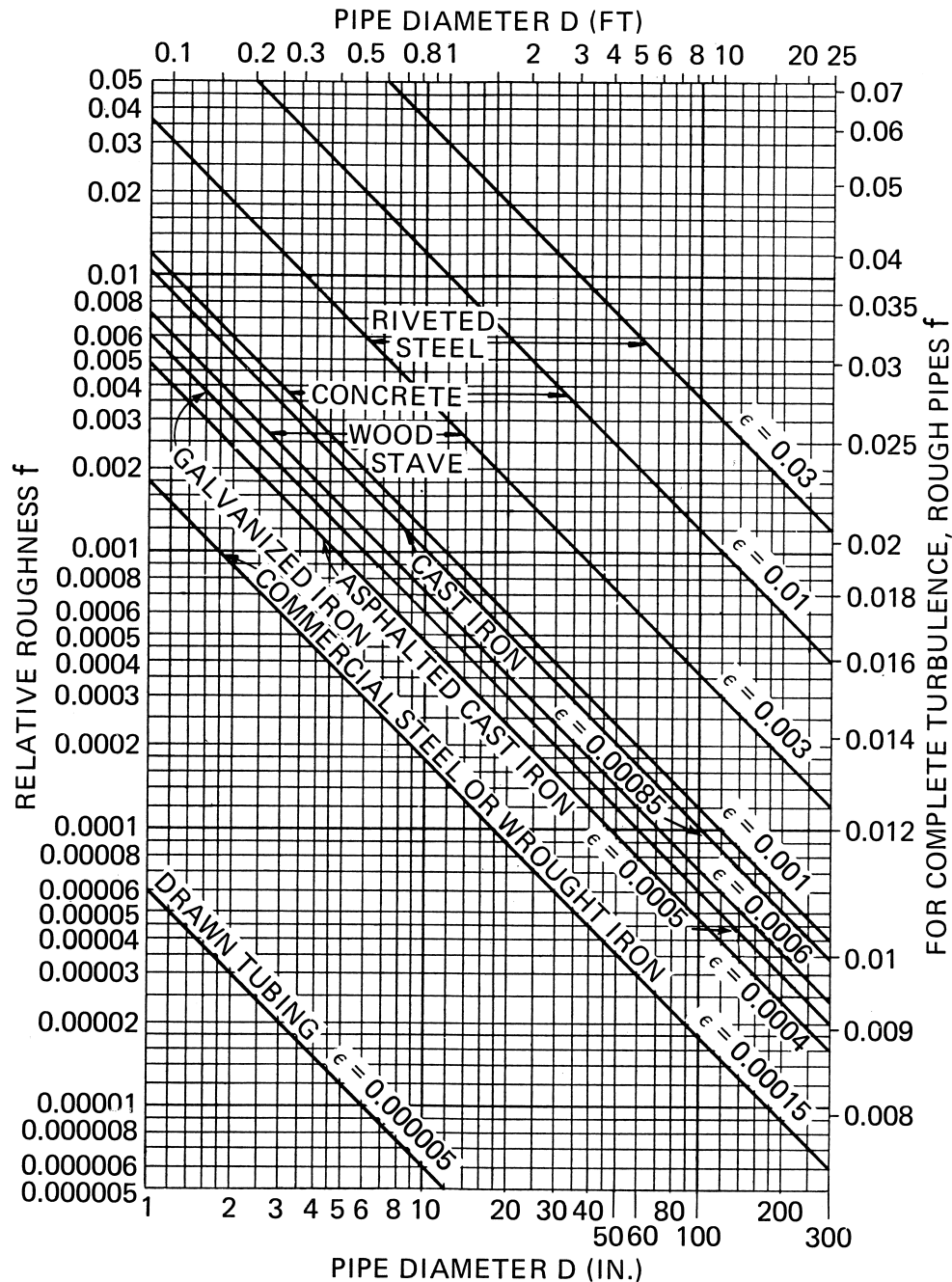


FIGURE 3B Roughness factors for selected types of pipe materials. [Adapted from Moody, L. F. (1944). *Trans. ASME* 66, 671–684.]

equivalent of the energy input to the fluid by a pump in the system. This is the actual or hydraulic head. In order to obtain the total work that a pump–motor combination performs, one must take into account the efficiency of the pump–motor set.

The efficiency is the ratio of the useful energy delivered to the fluid as work to the total energy consumed by the

motor and is always less than unity. This is a figure of merit of the pump that must be determined by experimentation and is supplied by the pump manufacturer. As a rule of thumb, well-designed centrifugal pumps usually operate at about 75–80% efficiency, while well-designed positive displacement pumps generally operate in the 90–95% efficiency range. In the case of positive displacement pumps,

the efficiency is determined primarily by the mechanical precision of the moving parts and the motor's electrical efficiency. Centrifugal pumps also depend strongly on the hydraulic conditions inside the pump and are much more variable in efficiency. More is said about this in Section VI.

d. Hydraulic grade line. Equation (63) is a finite difference equation and applies only to differences in the various energy quantities at two discrete points in the system. It takes no account of any conditions intermediate to these two reference points. If one were to keep point 1 fixed and systematically vary point 2 along the length of the pipe, the values of the various heads calculated would represent the systematic variation of velocity, potential, pressure, pump, and friction head along the pipe route. If all these values were plotted as a function of L , the distance down the pipe from point 1, a plot similar to that shown schematically in Fig. 4, would be obtained.

Figure 4 graphically illustrates the relation between the various heads in Eq. (63). For a pipe of constant cross section, the equation of continuity requires $\Delta(v)^2/2g = 0$. The pump, located as shown, creates a positive head $h_s = \hat{w}_s/g$, represented by the vertical line of this height at the pump station (PS). The straight line of slope $-h_f/L$ drawn through the point $(h_s, 0)$ is a locus of all values of potential (Δz), pressure ($\Delta p/\rho g$), and friction (h_f) heads calculated from Eq. (63) for any length of pipe (L). It is called the hydraulic grade line (HGL). The vertical distance between the HGL and the constant value $-h_s$ represents the energy that has been lost to that point due to friction. The height Δz designated as ground profile (GP) is a locus of physical ground elevations along the pipeline route and is also the actual physical location of the pipe itself. The difference between the HGL and the GP is the pressure head $\Delta p/\rho g$ at the length L . The significance of this head is that if one were to poke a hole in the pipe at point L , a fluid jet would spurt upward to a height equal to the HGL at that point. Thus, the HGL shows graphically at each point along the pipeline route the available pressure head to drive the flow through the pipe.

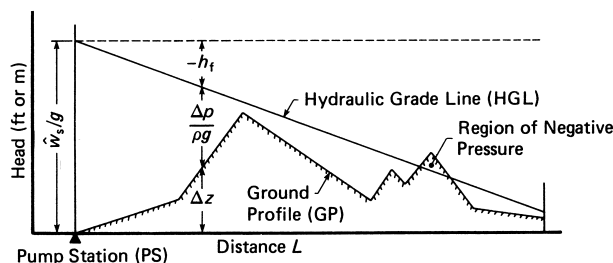


FIGURE 4 Illustration of hydraulic grade line concept.

If the HGL intersects and drops below the GP, as in the area of Fig. 4 marked “region of negative pressure,” there is not sufficient pressure head in the pipe to provide the potential energy necessary to raise the fluid to the height Δz at that point. Thus, if a hole were poked into the pipe at such a point, rather than a jet spurting out of the pipe, air would be drawn into the pipe. In a closed pipe a negative gauge pressure develops. This negative gauge pressure is the source of operation of a siphon. If, however, the absolute pressure in this part of the pipe drops to the vapor pressure of the liquid, the liquid will boil. This may cause the formation of a vapor bubble at the top of the pipe, or it may result in full vapor locking of the pipe, depending on the pressure conditions. This is called cavitation. Downhill of the negative pressure region where the HGL reemerges above the GP, the pressure rises back above the vapor pressure of the liquid and the vapor recondenses. This can occur with almost explosive violence and can result in physical damage to the pipe. Regions where this cavitation occurs are called “slack flow” regions. The HGL plot provides a simple and easy way to identify potential slack flow regions. In good design, such regions are avoided by the expedient of introducing another pump just upstream of the point where the HGL intersects the GP. Details of such procedures are discussed in Section VI.

3. Thermal Energy (Heat Transfer)

Just as the macroscopic mechanical energy equation is used to determine the relations between the various forms of mechanical energy and the frictional energy losses, so the thermal energy equation, expressed in macroscopic form, is used to determine the relation between the temperature and heat transfer rates for a flow system.

When Eq. (46) is applied to the thermal energy terms, we obtain

$$\frac{\partial U}{\partial t} + \langle \rho u \mathbf{v} \cdot \mathbf{n} \rangle_1 A_1 + \langle \rho u \mathbf{v} \cdot \mathbf{n} \rangle_2 A_2 = \dot{Q} + \dot{F} + \dot{Q}'_{CR}, \quad (67)$$

where U is the total internal energy of the fluid and the other terms have the significance already discussed.

Equation (67) is the basis of practical heat transfer calculations. In order to use it to solve problems, additional information is required about the total heat transfer rate $\dot{Q}$ and the production rate $\dot{Q}'_{CR}$. The first is usually expressed in terms of a heat transfer coefficient analogous to the friction factor,

$$\dot{Q} = U_m A_s \Delta T_m, \quad (68)$$

where U_m is an “overall” heat transfer coefficient, which is usually related to “local” heat transfer coefficients both

inside and outside the pipe; A_s is the area of the heated pipe surface; and ΔT_m is some sort of mean or average temperature difference between the fluid and the pipe wall. Depending on the definition of ΔT_m , the definitions of the local heat transfer coefficients vary and so does the definition of U_m . This equation is not discussed further here, as its full discussion properly belongs in a separate article devoted to the subject of heat transfer.

IV. LAMINAR FLOW

In laminar flow the velocity distribution, and hence the frictional energy loss, is governed entirely by the rheological constitutive relation of the fluid. In some cases it is possible to derive theoretical expressions for the friction factor. Where this is possible, a three-step procedure must be followed.

1. Solve the equations of motion for the stress distribution.
2. Couple the stress distribution with the constitutive relation to produce a differential equation for the velocity field. Solve this equation for the velocity distribution.
3. Integrate the velocity distribution over the cross section of the duct to obtain an expression for the average velocity $\langle v \rangle$. Rearrange this expression into a dimensionless form involving a friction factor.

A. Shear Stress Distributions

In some special cases it is possible to solve the equations of motion [Eq. (11)] entirely independently of any knowledge of the constitutive relation and to obtain a universal shear stress distribution that applies to all fluids. In other cases it is not possible to do this because the evaluation of certain integration constants requires knowledge of the specific constitutive relation. Because of space limitations, we illustrate only one case of each type here.

1. Pipes

Equation (11) for the cylindrical geometry appropriate to the circular cross-section pipes so commonly used in practical situations is expressed by Eqs. (17)–(19). For steady, fully developed, incompressible flow, the solution of these equations is

$$\tau_{rz} = \frac{r}{2} \left(-\frac{dp}{dz} \right) + \frac{C}{r}, \quad (69)$$

where C is a constant of integration. Considerations of boundedness at the pipe centerline, $r=0$, require that $C=0$. Thus, Eq. (69) reduces to the familiar linear stress distribution,

$$\tau_{rz} = \frac{r}{2} \left(-\frac{dp}{dz} \right) = \xi \tau_w, \quad (70)$$

where $-dp/dz$ is the axial pressure gradient, $\xi = r/R$ is a normalized radial position variable, and τ_w is the wall shear stress given by

$$\tau_w = \frac{D}{4} \left(-\frac{\Delta p}{L} \right). \quad (71)$$

Equation (70) is clearly independent of any constitutive relation and applies universally to all fluids in a pipe of this geometry.

2. Concentric Annulus

Suppose a solid core were placed along the centerline of the pipe described in the preceding section so as to be coaxial and concentric with the pipe. Equation (69) is still valid as the solution of Eqs. (17)–(19). Now, however, the point $r=0$ is not included in the domain of the solution, so that C is no longer zero. Somewhere between the two boundaries $r=R_i$ and $r=R$ the shear stress will vanish. If this point is called $\xi = \lambda$, then Eq. (69) becomes

$$\tau_{rz} = \tau_R (\xi - \lambda^2/\xi), \quad (72)$$

where τ_R is the shear stress at the outer pipe wall given by

$$\tau_R = \frac{R}{2} \left(-\frac{dp}{dz} \right) \quad (73)$$

and $\xi = r/R$ as before. Note two things: (1) Eq. (72) is now nonlinear in ξ , and (2) we still do not know the value of C . All that has been done is to shift the unknown value of C to the still unknown value of λ . We do, however, know the physical significance of λ . It is the location of the zero-stress surface. Unfortunately, we cannot discover the value of λ until we introduce some specific constitutive relation, integrate the resulting differential equation for the velocity distribution (thus introducing yet another constant of integration), and then invoke the no-slip or zero-velocity boundary conditions at *both* solid boundaries to determine the values of the new integration constant and λ . The value of λ so determined will be different for each different constitutive relation employed.

B. Velocity Distributions

1. Newtonian

When the Newtonian constitutive relation is coupled with Eq. (70) and appropriate integrations are performed, we obtain

$$u = v_z/\langle v \rangle = 2(1 - \xi^2) \quad (74)$$

$$\langle v \rangle = D\tau_w/8\mu \quad (75)$$

which are respectively known as the Poiseuille velocity profile and the Hagen–Poiseuille relation.

When the same operations are performed for the concentric annulus geometry, the results are

$$u = \frac{2}{F(\sigma)} [1 - \xi^2 + (1 - \sigma^2) \ln \xi / \ln(1/\sigma)] \quad (76)$$

$$\langle v \rangle = D\tau_R F(\sigma) / 8\mu \quad (77)$$

$$F(\sigma) = 1 + \sigma^2 - (1 - \sigma^2) / \ln(1/\sigma) \quad (78)$$

where $\sigma = R_i/R$ is the “aspect” ratio of the annulus.

2. Non-Newtonian

Because of the extreme complexity of the expressions for the velocity distributions and average velocities in concentric annuli for even simple non-Newtonian fluids, we include here only the results for pipe flow.

a. Bingham Plastic. The pertinent results are

$$u = \frac{2}{F(\xi_0)} [1 - \xi^2 - 2\xi_0(1 - \xi)], \quad \xi > \xi_0 \quad (79)$$

$$u = 2(1 - \xi_0)^2 / F(\xi_0), \quad \xi \leq \xi_0 \quad (80)$$

$$\langle v \rangle = D\tau_w F(\xi_0) / 8\mu_\infty \quad (81)$$

$$F(\xi_0) = 1 - \frac{4}{3}\xi_0 + \frac{1}{3}\xi_0^4 \quad (82)$$

where $\xi_0 = \tau_0/\tau_w$. Equation (81) is a version of the well-known Buckingham relation and is the Bingham plastic equivalent of the Hagen–Poiseuille result. The parameter ξ_0 , because of the linearity of Eq. (70), also represents the dimensionless radius of a “plug” or “core” of unsheared material in the center of the pipe, which moves at the maximum velocity given by Eq. (80). This is a feature of all fluids that possess yield stresses.

b. Power law. The pertinent results are

$$u = \frac{1+3n}{1+n} (1 - \xi^{(1+n)/n}) \quad (83)$$

$$\langle v \rangle = \frac{D}{2} \left(\frac{n}{1+3n} \right) \left(\frac{\tau_w}{k} \right)^{1/n} \quad (84)$$

Note that these results reduce to the Newtonian results in the limit $n = 1$, $k = \mu$.

c. Herschel–Bulkley. The pertinent results are

$$u = \frac{1+3n}{1+n} \frac{1}{F(\xi_0, n)} \left[(1 - \xi_0)^{(1+n)/n} - (\xi - \xi_0)^{(1+n)/n} \right], \quad \xi > \xi_0 \quad (85)$$

$$u = \frac{1+3n}{1+n} \frac{1}{F(\xi_0, n)} (1 - \xi_0)^{(1+n)/n}, \quad \xi \leq \xi_0 \quad (86)$$

$$\langle v \rangle = \frac{D}{2} \frac{n}{1+3n} \left(\frac{\tau_w}{k} \right)^{1/n} (1 - \xi_0)^{(1+n)/n} F(\xi_0, n) \quad (87)$$

$$F(\xi_0, n) = (1 - \xi_0)^2 + \frac{2(1+3n)\xi_0(1 - \xi_0)}{1+2n} + \frac{1+3n}{1+n} \xi_0^2 \quad (88)$$

where ξ_0 has the same significance as in the Bingham case.

d. Casson. The pertinent results are

$$u = \frac{2}{G(\xi_0)} \left[1 - \xi^2 + 2\xi_0(1 - \xi) - \frac{8}{3}\xi_0^{1/2}(1 - \xi^{3/2}) \right], \quad \xi > \xi_0 \quad (89)$$

$$u = \frac{2}{G(\xi_0)} \left(1 - \frac{8}{3}\xi_0^{1/2} + 2\xi_0 - \frac{1}{3}\xi_0^2 \right), \quad \xi \leq \xi_0 \quad (90)$$

$$\langle v \rangle = D\tau_w G(\xi_0) / 8\mu_\infty \quad (91)$$

$$G(\xi_0) = 1 - \frac{16}{7}\xi_0^{1/2} + \frac{4}{3}\xi_0 - \frac{1}{21}\xi_0^4 \quad (92)$$

where ξ_0 has the same significance as in the Bingham case.

It should be observed that in all cases, even the *linear* Bingham plastic case, the resultant average velocity expressions are nonlinear relations between $\langle v \rangle$ and $-dp/dz$. This is true of all non-Newtonian constitutive relations. A direct consequence of this result is that the friction factor relation is also nonlinear.

C. Friction Factors

In Eq. (65) the friction factor was introduced as an empirical factor of proportionality in the calculation of the friction loss head. If Eq. (63) is applied to a length of straight horizontal pipe with no pumps, one finds that

$$-h_f = \Delta p / \rho g. \quad (93)$$

Elimination of h_f between Eqs. (65) and (93) results in

$$f = \frac{8}{\rho \langle v \rangle^2} \left(\frac{-D \Delta p}{4L} \right) = \frac{8\tau_w}{\rho \langle v \rangle^2}, \quad (94)$$

which may be looked on as an alternate definition of the friction factor. From Eq. (66) it is evident that Eq. (94) with the numeric factor 8 replaced by 2 defines the Fanning friction factor.

1. Newtonian

Equation (94) provides the means for rearranging all of the theoretical expressions for $\langle v \rangle$ given above into expressions involving the friction factor. For example, when Eq. (75) for Newtonian pipe flow is so rearranged and one eliminates $\langle v \rangle$ in terms of the Reynolds number, $\text{Re} = D\langle v \rangle \rho / \mu$, one obtains

$$f = 64/\text{Re}. \quad (95)$$

Equation (95) is the source of the laminar flow line on the Moody chart (Fig. 3).

In the case of the concentric annulus the problem is somewhat ambiguous, because there are two surfaces of different diameter and hence the specification of a length in Re is not obvious as in the case of the pipe. For example, one could use D_i , D , or $D - D_i$ or a host of other possibilities. Obviously, for each choice a different definition of Re arises. Also, the specification of τ_w in Eq. (94) is ambiguous for the same reason. Here, we list only one of many possible relations,

$$f'_R = 2\tau_R / \rho \langle v \rangle^2 = 16/F(\sigma)\text{Re}_D, \quad (96)$$

where both f'_R and Re_D are based on τ_w and D for the outer pipe. The function $F(\sigma)$ in Eq. (96) is the same as given by Eq. (78).

2. Non-Newtonian

The ambiguity of definition of Re encountered in the concentric annulus case is compounded here because of the fact that no “viscosity” is definable for non-Newtonian fluids. Thus, in the literature one encounters a bewildering array of definitions of Re -like parameters. We now present friction factor results for the non-Newtonian constitutive relations used above that are common and consistent. Many others are possible.

a. Bingham plastic. The pertinent results are

$$f' = \frac{16}{\text{Re}_{\text{BP}}} + \frac{8}{3} \frac{\text{He}}{\text{Re}_{\text{BP}}^2} - \frac{16}{3} \frac{\text{He}^4}{f'^3 \text{Re}_{\text{BP}}^8} \quad (97)$$

$$\text{He} = D^2 \rho \tau_0 / \mu_\infty^2 \quad (98)$$

$$\text{Re}_{\text{BP}} = D\langle v \rangle \rho / \mu_\infty \quad (99)$$

Note that a new dimensionless parameter He , called the Hedstrom number, arises because in the constitutive relation there are two independent rheological parameters. Parameter He is essentially a dimensionless τ_0 . This multiplicity of dimensionless parameters in addition to the Re parameter is common to all non-Newtonian constitutive relations.

b. Power law. The pertinent results are

$$f' = 16/\text{Re}_{\text{PL}} \quad (100)$$

$$\text{Re}_{\text{PL}} = 2^{3-n} \frac{D^n \langle v \rangle^{2-n} \rho}{k} \left(\frac{n}{1+3n} \right)^n \quad (101)$$

Historically, Re_{PL} was invented to force the form of Eq. (100).

c. Herschel–Bulkley. The pertinent results are

$$f' = 16 / [\text{Re}_{\text{HB}}(1 - \xi_0)^{1+n} F(\xi_0, n)^n] \quad (102)$$

$$\xi_0 = \frac{2}{f'} \left[\frac{\text{He}_{\text{HB}}^n \left(\frac{n}{1+3n} \right)^{2n} (2^{3-n})^2}{\text{Re}_{\text{HB}}^2} \right]^{1/(2-n)} \quad (103)$$

$$\text{He}_{\text{HB}} = \frac{D^2 \rho}{\tau_0} (\tau_0/k)^{2/n} \quad (104)$$

and Re_{HB} is identical in definition to Eq. (101). Indeed, Eqs. (102)–(104) reduce to Eqs. (100) and (101) for the limit $\tau_0 = 0$. In Eq. (104) He_{HB} is the Herschel–Bulkley equivalent of the Bingham plastic Hedstrom number He .

d. Casson. The pertinent results are

$$f' = 16/\text{Re}_{\text{CA}} G'(f', \text{Ca}, \text{Re}_{\text{CA}}) \quad (105)$$

$$G'(f', \text{Ca}, \text{Re}_{\text{CA}}) = 1 - \frac{16\sqrt{2}}{7} \frac{(\text{Ca}/f')^{1/2}}{\text{Re}_{\text{CA}}} + \frac{8}{3} \frac{(\text{Ca}/f')}{\text{Re}_{\text{CA}}^2} - \frac{16(\text{Ca}/f')^4}{21 \text{Re}_{\text{CA}}^8} \quad (106)$$

$$\text{Ca} = D^2 \rho \tau_0 / \mu_\infty^2 \quad (107)$$

$$\text{Re}_{\text{CA}} = D\langle v \rangle \rho / \mu_\infty \quad (108)$$

The parameter Ca is called the Casson number and is analogous to the Hedstrom number He for the Bingham plastic and Herschel–Bulkley models.

V. TURBULENT FLOW

A. Transition to Turbulence

As velocity of flow increases, a condition is eventually reached at which rectilinear laminar flow is no longer stable, and a transition occurs to an alternate mode of motion that always involves complex particle paths. This motion may be of a multidimensional secondary laminar form, or it may be a chaotic eddy motion called turbulence. The nature of the motion is governed by both the rheological nature of the fluid and the geometry of the flow boundaries.

1. Newtonian

The most important case of this transition for chemical engineers is the transition from laminar to turbulent flow, which occurs in straight bounded ducts. In the case of Newtonian fluid rheology, this occurs in straight pipes when $Re = 2100$. A similar phenomenon occurs in pipes of other cross sections, as well and also for non-Newtonian fluids. However, just as the friction factor relations for these other cases are more complex than for simple Newtonian pipe flow, so the criteria for transition to turbulence cannot be expressed as a simple critical value of a Reynolds number.

All pressure-driven, rectilinear duct flows, whether Newtonian or non-Newtonian, undergo transition to turbulence when the transition parameter K_H of Hanks, defined by

$$K_H = \frac{\rho |\nabla v^2|/2}{|\rho \mathbf{g} - \nabla p|}, \quad (109)$$

achieves a maximum value of 404 at some point in the duct flow. In this equation v is the laminar velocity distribution. In the special limit of Newtonian pipe flow, Eq. (109) reduces the $Re_c = 2100$. For the concentric annulus, it reduces to

$$Re_{DC} = 808F(\sigma)/[(1 - \bar{\xi}^2 + 2\lambda^2 \ln \bar{\xi})|\bar{\xi} - \lambda^2/\bar{\xi}|], \quad (110)$$

where $\bar{\xi}$ is the root of

$$(1 - \bar{\xi}^2 + 2\lambda^2 \ln \bar{\xi})(\lambda^2 + \bar{\xi}^2) - 2(\bar{\xi}^2 - \lambda^2)^2 = 0 \quad (111)$$

with λ defined by

$$\lambda^2 = \frac{1}{2}(1 - \sigma^2)/\ln(1/\sigma) \quad (112)$$

and $F(\sigma)$ is given by Eq. (78). There are two roots to Eq. (111), with the result that Eq. (110) predicts two distinct Reynolds numbers of transition, in agreement with experiment.

2. Non-Newtonian

a. Bingham plastic. The critical value of Re_{BP} is given by

$$Re_{BPc} = He(1 - \frac{4}{3}\xi_{0c} + \frac{1}{3}\xi_{0c}^4)/8\xi_{0c}, \quad (113)$$

where He is the Hedstrom number and ξ_{0c} is the root of

$$\xi_{0c}/(1 - \xi_{0c})^3 = He/16,800. \quad (114)$$

The predictions of these equations agree very well with experimental data.

b. Power law. The pertinent results are

$$Re_{PLc} = \frac{6464n}{(1 + 3n)^2}(2 + n)^{(2+n)/(1+n)}. \quad (115)$$

c. Herschel–Bulkley. The pertinent results are

$$Re_{HBC} = \frac{6464n}{(1 + 3n)^2}(2 + n)^{(2+n)/(1+n)} \left[\frac{F(\xi_{0c}, n)^{2-n}}{(1 - \xi_{0c})^n} \right], \quad (116)$$

where ξ_{0c} is the root of

$$\left[\frac{\xi_{0c}}{(1 - \xi_{0c})^{1+n}} \right]^{(2-n)/n} \left[\frac{1}{(1 - \xi_{0c})^n} \right] = \frac{nHe_{HB}}{3232(2 + n)^{(2+n)/(1+n)}}. \quad (117)$$

He_{HB} is given by Eq. (104) and $F(\xi_{0c}, n)$ is given by Eq. (88) evaluated with $\xi = \xi_{0c}$.

d. Casson. The pertinent equations are

$$Re_{CAC} = CaG(\xi_{0c})/8\xi_{0c}, \quad (118)$$

where ξ_{0c} must be determined from the simultaneous solution of Eqs. (119) and (120),

$$0 = 1 + 2\xi_{0c} - \frac{8}{3}\xi_{0c}^{1/2} + 2\bar{\xi}^{1/2}\xi_{0c}^{3/2} - 8\xi_{0c}\bar{\xi} + \frac{26}{3}\xi_{0c}^{1/2}\bar{\xi}^{3/2} - 3\bar{\xi}^2 \quad (119)$$

$$6464\xi_{0c}/Ca = [1 - \bar{\xi}^2 + 2\xi_{0c}(1 - \bar{\xi}) - \frac{8}{3}\xi_{0c}^{1/2}(1 - \bar{\xi}^{3/2})](\bar{\xi}^{1/2} - \xi_{0c}^{1/2})^2 \quad (120)$$

and $G(\xi_{0c})$ is given by Eq. (92) evaluated with $\xi = \xi_{0c}$.

B. Reynolds Stresses

When full turbulence occurs, the details of the velocity distribution become extremely complicated. While in principle these details could be computed by solving the general field equations given earlier, in practice it is essentially impossible. As an alternative to direct solution it is customary to develop a new set of equations in terms of Reynolds' averages. The model is illustrated schematically in Fig. 5.

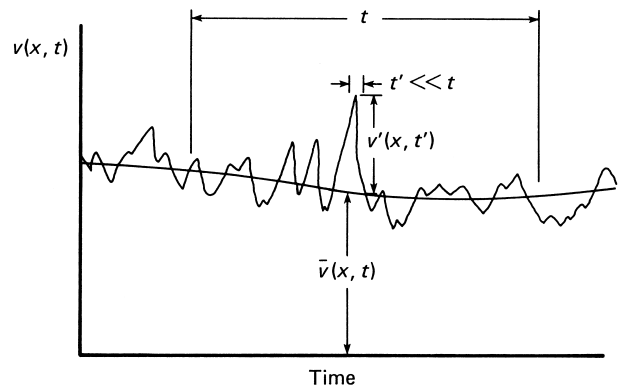


FIGURE 5 Schematic illustration of Reynolds' convention $\mathbf{v} = \bar{\mathbf{v}} + \mathbf{v}'$ for turbulent flow.

The actual velocity field fluctuates wildly. Reynolds modeled it by a superposition of a Eulerian time mean value $\bar{\mathbf{v}}$ defined by

$$\bar{\mathbf{v}}(\mathbf{x}, t) = \frac{1}{t} \int_0^t \mathbf{v}(\mathbf{x}, t') dt', \quad (121)$$

where t' is a time interval of the order of an individual excursion and t is a time interval large in comparison with t' but small enough that gross time variations of the mean field can still be observed and calculated by the basic field equations. In terms of this model then, we write

$$\mathbf{v} = \bar{\mathbf{v}} + \mathbf{v}' \quad (122)$$

with $\mathbf{v}'$ being the instantaneous excursion or “fluctuation” from $\bar{\mathbf{v}}$. After this result is introduced into the field equations and the time-averaging operation defined in Eq. (121) is invoked, we obtain a new set of averaged field equations for the turbulent flow.

For incompressible fluids we obtain the following results:

Equation of Continuity

$$\nabla \cdot \bar{\mathbf{v}} = 0 \quad (123)$$

Cauchy's Equations of Motion

$$\rho \frac{D\bar{\mathbf{v}}}{Dt} + \nabla \cdot \rho \overline{\mathbf{v}'\mathbf{v}'} = \rho \mathbf{g} - \nabla \bar{p} - \nabla \cdot \bar{\boldsymbol{\tau}} \quad (124)$$

Thermal Energy Relation

$$\begin{aligned} \rho \frac{\partial \bar{u}}{\partial t} + \rho \bar{\mathbf{v}} \cdot \nabla \bar{u} + \rho \overline{\mathbf{v}' \cdot \nabla u'} \\ = -\bar{\boldsymbol{\tau}} : \nabla \bar{\mathbf{v}} - \overline{\boldsymbol{\tau}' : \nabla \mathbf{v}'} - \nabla \cdot \bar{\mathbf{q}} + \bar{r}_{CR} \end{aligned} \quad (125)$$

Mechanical Energy Relation

$$\begin{aligned} \rho(\partial/\partial t)(\bar{v}^2/2) + \rho(\partial/\partial t)\overline{(v'^2/2)} + \rho \bar{\mathbf{v}}\bar{\mathbf{v}} : \nabla \bar{\mathbf{v}} \\ + \rho \overline{\mathbf{v}'\mathbf{v}' : \nabla \mathbf{v}'} + \rho \overline{\mathbf{v}'\bar{\mathbf{v}} : \nabla \mathbf{v}'} + \rho \overline{\mathbf{v}'\mathbf{v}' : \nabla \bar{\mathbf{v}}} \\ + \rho \overline{\mathbf{v}'\mathbf{v}' : \nabla p'} = \rho \bar{\mathbf{v}} \cdot \mathbf{g} - \bar{\mathbf{v}} \cdot \nabla \bar{p} \\ - \overline{\mathbf{v}' \cdot \nabla p'} - \bar{\mathbf{v}} \cdot (\nabla \cdot \bar{\boldsymbol{\tau}}) - \overline{\mathbf{v}' \cdot (\nabla \cdot \boldsymbol{\tau}')} \end{aligned} \quad (126)$$

Entropy Production Postulate

$$-\bar{\boldsymbol{\tau}} : \nabla \bar{\mathbf{v}} - \overline{\boldsymbol{\tau}' : \nabla \mathbf{v}'} \geq 0 \quad (127)$$

All of these relations contains terms involving statistical correlations among various products of fluctuating velocity, pressure, and stress terms. This renders them considerably more complex than their laminar flow counterparts. Reynolds succeeded in partially solving this dilemma by the expedient of introducing the turbulent stress tensor $\hat{\boldsymbol{\tau}}$, defined by

$$\hat{\boldsymbol{\tau}} = \bar{\boldsymbol{\tau}} + \rho \overline{\mathbf{v}'\mathbf{v}'}. \quad (128)$$

With this substitution Eq. (124) becomes identical with Eq. (10), with all terms replaced by their Eulerian time

mean values. Thus, any solution of Eq. (10) for the stress distribution also becomes a solution of Eq. (124) for the “turbulent” stress distribution. However, this small success extracts a dear price. No further progress can be made because a new unknown quantity, $\rho \overline{\mathbf{v}'\mathbf{v}'}$, which has come to be known as the Reynolds’ stress tensor, has been introduced with no compensating new equation for its calculation. This is the famous turbulence “closure” problem.

An enormous amount of effort has been expended in attempting to discover new equations for $\rho \overline{\mathbf{v}'\mathbf{v}'}$. Five different levels of approach have been pursued in the literature involving various degrees of mathematical complexity. We cannot discuss all of them here. We outline only two of the most fruitful: (1) the mixing length or zero-equation models and (2) the κ - ε or two-equation models.

C. Mixing Length Models

An early approach to the closure, typified by the work of Prandtl, represented the Reynolds’ stress tensor as

$$\rho \overline{\mathbf{v}'\mathbf{v}'} = 2\rho \hat{\epsilon}_\tau \cdot \bar{\mathbf{D}}, \quad (129)$$

where $\hat{\epsilon}_\tau$ is a second-order eddy diffusivity tensor and $\bar{\mathbf{D}}$ is the symmetric part of $\nabla \bar{\mathbf{v}}$ defined by Eq. (37) for $\mathbf{v} = \bar{\mathbf{v}}$. In this degree of approximation $\hat{\epsilon}_\tau$ is assumed to depend only on the properties of the mean velocity gradient tensor $\bar{\mathbf{D}}$ and is

$$\hat{\epsilon}_\tau = 2L^2 |\sqrt{-2II_{\bar{\mathbf{D}}}}| \delta, \quad (130)$$

where L is some sort of length measure of the turbulence called a mixing length, and $II_{\bar{\mathbf{D}}}$ is defined by Eq. (38) for $\mathbf{v} = \bar{\mathbf{v}}$.

For the special case of pipe flow, Prandtl modeled L as

$$L = k_t R(1 - \xi), \quad (131)$$

where k_t , known as the Von Karman constant, is an empirical parameter usually taken to be ~ 0.36 . This simple model leads to a rather famous results for the velocity distribution in a pipe:

$$u^+ = v/v^* = \frac{1}{0.36} \ln y^+ + 3.80, \quad y^+ > 26 \quad (132)$$

$$y^+ = \frac{Rv^*\rho}{\mu}(1 - \xi) \quad (133)$$

$$v^* = \sqrt{\tau_w/\rho} \quad (134)$$

The dimensionless variables u^+ and y^+ are called Prandtl’s universal velocity profile variables. The parameter v^* is called the friction velocity.

In efforts to increase the range of applicability of the mixing length model, numerous others have modified it.

One of the better versions of the modified mixing length model is

$$L = k_t R(1 - \xi)(1 - E) \quad (135)$$

$$E = \exp[-\phi^*(1 - \xi)] \quad (136)$$

$$\phi^* = (R^* - R_c^*)/2\sqrt{2}B \quad (137)$$

$$R^* = \text{Re}\sqrt{f'} \quad (138)$$

The parameter R_c^* is the laminar–turbulent transition value of R^* and has the numerical value 183.3 for Newtonian fluids. For non-Newtonian fluids it would have to be computed from the various results presented above.

The parameter B is called a dampening parameter, as its physical significance is associated with dampening turbulent fluctuations in the vicinity of a wall. For Newtonian pipe flow it has the numerical value 22. For non-Newtonian fluids it has been found to be a function of various rheological parameters as follows:

Bingham Plastic

$$B_{BP} = 22[1 + 0.00352\text{He}/(1 + 0.000504\text{He})^2] \quad (139)$$

Power Law

$$B_{PL} = 22/n \quad (140)$$

Herschel–Bulkley

$$B_{HB} = B_{BP}/n \quad (141)$$

No correlation has as yet been developed for the Casson model.

These simple models of turbulent pipe flow for various rheological models do not produce accurate details regarding the structure of the turbulent flow. They do, however, offer the practicing design engineer the opportunity to predict the gross engineering characteristics of interest with reasonable correctness. They are called zero-order equations because no differential equations for the turbulence properties themselves are involved in their solutions. Rather, one must specify some empirical model, such as the mixing length, to close the equations.

D. Other Closure Models

Actually, all methods of closure involve some type of modeling with the introduction of adjustable parameters that must be fixed by comparison with data. The only question is where in the hierarchy of equations the empiricism should be introduced. Many different systems of modeling have been developed. The zero-equation models have already been introduced. In addition there are one-equation and two-equation models, stress-equation models, three-equation models, and large-eddy simulation models. Depending on the complexity of the model and the problem

being investigated, one can obtain various degrees of detailed information about the turbulent motions. Most of the more complex formulations require large computing facilities and may result in extreme numerical stability and convergence problems. All of the different methods of computing the turbulent field structure cannot be discussed here. Therefore, only one of these other methods, the so-called κ – ε method, which is a two-equation type of closure model, is outlined. Models such as this have to date been applied only to Newtonian flow problems.

The idea involved in the κ – ε model is to assume that the Reynolds' stress tensor can be written as

$$\rho \overline{\mathbf{v}'\mathbf{v}'} = 2\mu_t \bar{\mathbf{D}} - \frac{2}{3}\kappa \delta, \quad (142)$$

where κ is the turbulent kinetic energy,

$$\kappa = \frac{1}{2} \overline{\mathbf{v}' \cdot \mathbf{v}'}, \quad (143)$$

and μ_t is a turbulent or eddy viscosity function quite analogous to the eddy diffusivity discussed earlier. Just as in the zero-equation modeling situation, one cannot write down a general defining equation for μ_t , but must resort to modeling. In the present case the model used is

$$\mu_t = c_1 \rho \kappa^2 / \varepsilon, \quad (144)$$

where c_1 is a (possibly) Reynolds number-dependent coefficient that must be determined empirically. The function ε is the turbulent energy dissipation rate function. The functions κ and ε are determined by the pair of simultaneous differential equations

$$\frac{D\kappa}{Dt} = \nabla \cdot \left(c_2 \frac{\mu_t}{\rho} \nabla \kappa \right) + \overline{\tau' : \nabla \mathbf{v}'} - \varepsilon \quad (145)$$

$$\frac{D\varepsilon}{Dt} = \nabla \cdot \left(c_3 \frac{\mu_t}{\rho} \nabla \varepsilon \right) + c_4 \frac{\varepsilon}{\kappa} \overline{\tau' : \nabla \mathbf{v}'} - c_5 \varepsilon^2 / \kappa \quad (146)$$

In this model the coefficients c_1 to c_5 are commonly given the numerical values $c_1 = 0.09$, $c_2 = 1.0$, $c_3 = 0.769$, $c_4 = 1.44$, and $c_5 = 1.92$, although these values can be varied at will by the user and are definitely problem specific. They can also be made functions of any variables necessary.

Here ends this article's discussion of this model, but extensive detail is available in numerous books on the subject. Some of these models present very accurate, detailed descriptions of the turbulence in some cases, but may be very much in error in others. Considerable skill and experience are required for their use.

VI. APPLICATIONS

A. Friction Factors

From a practical point of view the chemical engineer is very often interested in obtaining a relation between the overall pressure drop across a pipe, fitting, or piece of processing equipment and the bulk or mean velocity of flow through it. On occasion the details of the velocity, temperature, or concentration profile are important, but most frequently it is the gross pressure drop-flow rate behavior that is important to a chemical engineer.

This is generally obtained by use of the integrated form of the mechanical energy equation with the frictional energy loss calculated by Eq. (65). Thus, the basic problem facing a design engineer is how to obtain numerical values for the friction factor f .

For Newtonian fluids this problem is solved empirically by the introduction of the Moody diagram (Fig. 3). In the case of non-Newtonian fluids, however, this is not appropriate and alternative, semi-theoretical formulations must be developed. The theoretical laminar flow equations for the four rheological models considered here have already been presented, as have the modified mixing length models for turbulent flow of three of these same models. The latter equations can be integrated to obtain velocity distributions, which can in turn be integrated to produce mean velocity-pressure gradient relations. These results can then be algebraically rearranged into the desired friction factor correlations. These results are presented in the following subsections.

1. Bingham Plastic Pipe Flow

When the appropriate integrations are performed using the Bingham model, one obtains

$$\text{Re}_{\text{BP}} = \frac{1}{2} R_{\text{BP}}^{*2} \int_{\xi_0}^1 \xi^2 g(\xi, \xi_0, R_{\text{BP}}^*) d\xi \quad (147)$$

$$g(\xi, \xi_0, R_{\text{BP}}^*) = \frac{\xi - \xi_0}{1 + \left[1 + \frac{1}{2} k_t^2 R_{\text{BP}}^{*2} (\xi - \xi_0)(1 - \xi)^2 (1 - E)^2\right]^{1/2}}, \quad (148)$$

where E and R_{BP}^* are defined by Eqs. (136) to (138), with Re being replaced by Re_{BP}^* and B being given by Eq. (139). The parameter ξ_0 is related to R_{BP}^* by the relation

$$R_{\text{BP}}^{*2} = 2\text{He}/\xi_0. \quad (149)$$

These equations can be used for practical calculations in two ways. One may generate the equivalent of the Moody

diagram from them, or one may solve them iteratively for a specific design case. Both methods are illustrated below.

a. General friction factor plot. The Hedstrom number is the key design parameter. From its definition in Eq. (98) it can be seen that He depends only on the rheological parameters and the pipe diameter. The rheological parameters are obtained from laboratory viscometry data, and the pipe diameter is at the discretion of the designer to specify. Thus, its numerical value is discretionary.

For a given value of He one can compute a complete f' - Re curve as follows:

1. Compute Re_{BPC} from Eqs. (113) and (114).
2. Using ξ_{0c} in Eq. (149) compute R_{BPC}^* .
3. For $\text{Re}_{\text{BP}} < \text{Re}_{\text{BPC}}$ compute f' from Eq. (97).
4. For $\text{Re}_{\text{BP}} > \text{Re}_{\text{BPC}}$ choose a sequence of values of $R_{\text{BP}}^* > R_{\text{BPC}}^*$.
5. For each such value of R_{BP}^* compute ξ_0 from Eq. (149) and Re_{BP} from Eqs. (147) and (148).
6. From the computed value of Re_{BP} and the assumed value of R_{BP}^* compute f' from Eq. (138).
7. Repeat steps 4–6 as many times as desired and plot the pairs of points f' , Re_{BP} so computed to create the equivalent Moody plot.

Figure 6 was created in this manner for a series of decade values of He . It may be used in place of the Moody chart for standard pipeline design problems. Because of the manner in which the empirical correlation for B was determined, no correction for pipe relative roughness is needed when one is dealing with commercial grade-steel line pipe.

b. Specific design conditions. A very common design situation involves the specification of a specific throughput and pipe diameter, thus fixing Re_{BP} but not R_{BP}^* . The system of equations presented earlier must therefore be solved iteratively for the value of R_{BP}^* , which produces the design Re_{BP} from Eq. (147).

The procedure to be followed is nearly the same as that already outlined. Steps 1 and 2 are followed exactly to determine R_{BPC}^* . Steps 4 and 5 are repeated iteratively until the Re_{BP} computed from Eq. (147) agrees with the design Re_{BP} to some acceptable convergence criterion. Because of the pinching effect of the curves in Fig. 6 at larger Re_{BP} values, it is best to use slower but more reliable interval halving techniques in searching for the root of the equation rather than a faster but often unstable Newton-Raphson method.

As an alternative to these two techniques, which involve considerable programming and numerical integration,

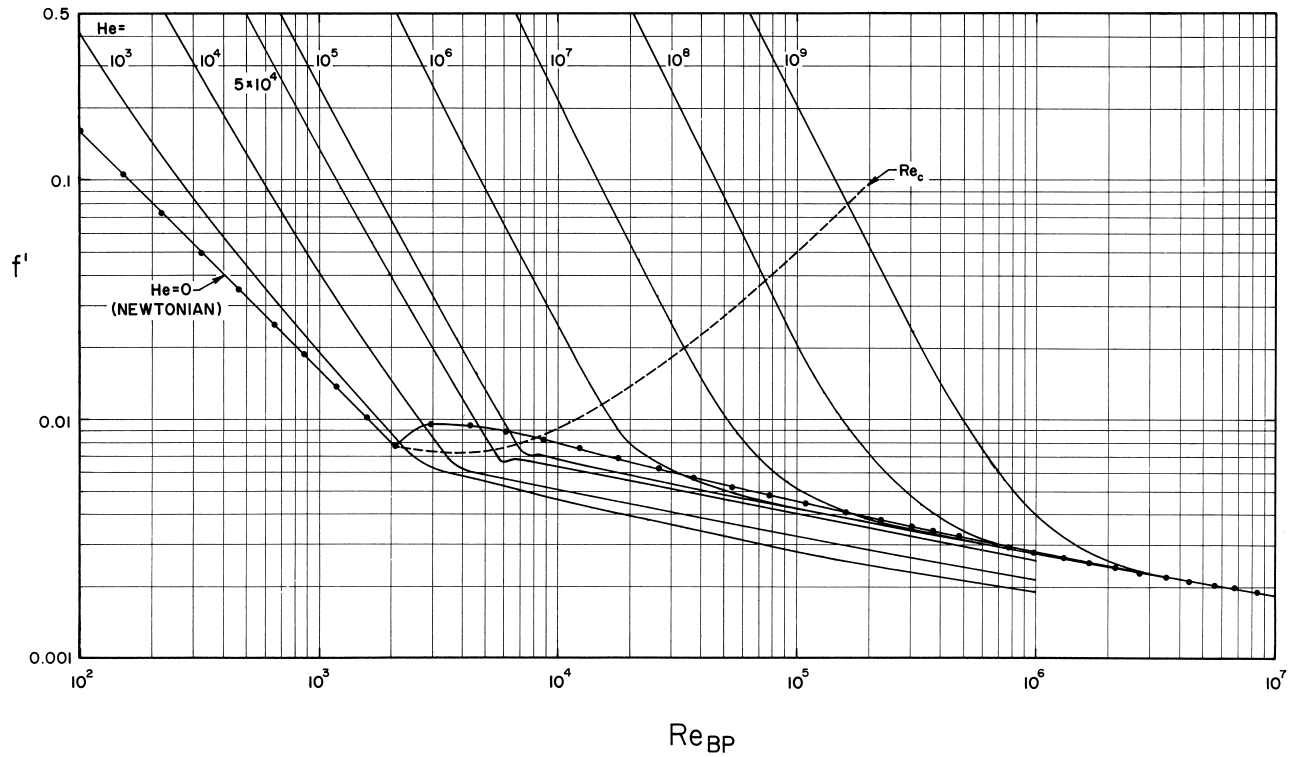


FIGURE 6 Fanning friction factor–Bingham plastic Reynolds number curves for Bingham plastic fluids. [Reproduced from Hanks, R. W. (1981). “Hydraulic Design from Flow of Complex Mixtures,” Richard W. Hanks Associates, Inc., Orem, UT.]

the following empirical curve fits of Fig. 6 have been developed:

$$f' = 10^A / \text{Re}_{\text{BP}}^{0.193} \quad (150)$$

$$A = -1.378\{1 + 0.146 \exp[-2.9(10^{-5})\text{He}]\} \quad (151)$$

These equations are valid only for turbulent flow.

2. Power Law Model Pipe Flow

The pertinent equations here are

$$\text{Re}_{\text{PL}} = \left(\frac{n}{1+3n} \right)^n R_{\text{PL}}^{*2} \left[\int_0^1 \xi^2 \zeta(\xi, R_{\text{PL}}^*) d\xi \right]^{2-n} \quad (152)$$

$$R_{\text{PL}}^* = \frac{3n+1}{n} \left[\text{Re}_{\text{PL}} \left(\frac{f'}{16} \right)^{(2-n)/2} \right]^{1/n} \quad (153)$$

$$\xi = \zeta^n + \frac{1}{8} R_{\text{PL}}^{*2} L_{\text{PL}}^{*2} \zeta^2 \quad (154)$$

As with the Bingham case one first computes Re_{PLC} from Eq. (115) and then uses Eq. (153) to compute R_{PLC}^* . The value of f' to be used in this calculation comes from Eq. (100). Once R_{PLC}^* is known, one then chooses a series of values of $R_{\text{PL}}^* > R_{\text{PLC}}^*$ and computes Re_{PL} for each

from Eq. (152). These values, together with the specified values of R_{PL}^* and Eq. (153), determine the corresponding values of f' . In Eq. (152) the function $\zeta(\xi, R_{\text{PL}}^*)$ is defined implicitly by Eq. (154), where the mixing length L_{PL}^* is equal to L_{PL}/R , with L_{PL} being determined by Eqs. (135)–(137) and (140). The computation of f' for a specific value of Re_{PL} is carried out iteratively using these equations in exactly the same manner as described for the Bingham model.

An approximate value of f' can be computed from the following empirical equation:

$$\sqrt{\frac{1}{f'}} = \frac{4.0}{n^{0.75}} \log(\text{Re}_{\text{PL}} f'^{(2-n)/2}) - \frac{0.4}{n^{1.2}}. \quad (155)$$

3. Herschel–Bulkley Model Pipe Flow

For this model the pertinent equations are

$$\text{Re}_{\text{HB}} = (1 - \xi_0)^{(2-n)/n} \left(\frac{n}{1+3n} \right)^n R_{\text{HB}}^{*2} \times \left[\int_{\xi_0}^1 \xi^2 \zeta(\xi, \xi_0, R_{\text{HB}}^*) d\xi \right]^{2-n} \quad (156)$$

$$\xi = \xi_0 + (1 - \xi_0)\zeta^n + \frac{1}{8}R_{HB}^{*2}(1 - \xi_0)^{2/n}L_{HB}^{*2}\zeta^2 \quad (157)$$

$$R_{HB}^{*2} = 2He_{HB}/\xi_0^{(2-n)/n} \quad (158)$$

Equation (156) is exactly analogous to Eq. (152) for the power law model and to Eq. (147) for the Bingham model. R_{HB}^* is defined in relation to Re_{HB} and f' by Eq. (153), with Re_{PL} being replaced by Re_{HB} . The function $\xi(\xi, \xi_0, R_{HB}^*)$ is defined implicitly by Eq. (157), with $L_{HB}^* = L_{HB}/R$, and L_{HB} is given by Eqs. (135)–(137) and (141). The value of ξ_0 to be used in all of these equations is determined from Eq. (158) for specified values of He and R_{HB}^* . The computational procedures follow exactly the steps outlined for the other models. There are no simple empirical expressions that can be used to bypass the numerical integrations called for by this theory. One must use the above equations.

4. Casson Model Pipe Flow

As of the time of this writing, the corresponding equations for the Casson model have been developed but have not been tested against experimental data. Therefore, we cannot include any results.

5. Other Non-Newtonian Fluids

Thus far we have given exclusive attention to the flow of purely viscous fluids. In practice the chemical engineer often encounters non-Newtonian fluids exhibiting elastic as well as viscous behavior. Such viscoelastic fluids can be extremely complex in their rheological response. The level of mathematical complexity associated with these types of fluids is much more sophisticated than that presented here. Within the limits of space allocated for this article, it is not feasible to attempt a summary of this very extensive field. The reader must seek information elsewhere. Here we shall content ourselves with fluids that do not exhibit elastic behavior.

B. Pipeline System Design

1. Hydraulic Grade Line Method

As already indicated, once one has in hand a method for estimating friction factors, the practical engineering problem of designing pumping systems rests on systematic application of the macroscopic or integrated form of the mechanical energy equation [Eq. (63)], with h_f being defined in terms of f by Eq. (65). Section III.C.2.d introduced the concept of the hydraulic grade line, or HGL. This is simply a graphic representation of the locus of all possible solutions of Eq. (63) along a given pipeline for a given flow rate. When coupled with a ground profile (GP) as illustrated schematically in Fig. 4, this plot provides a

particularly useful and simple means of identifying potential trouble spots in a pipeline. Although in this age of computers graphic techniques have generally fallen into disuse, this method still finds active use in commercial pipeline design practice.

The method is applied as illustrated below for a typical design problem. The conditions of the problem are $Q = 17,280$ bbl/day (528 gpm or 0.0333 m³/sec) of a Newtonian fluid of specific gravity = 1.18 and viscosity = 4.1 cP (0.0041 Pa · sec) with a reliability factor of 0.9 and a terminal end head of 100 ft (30.48 m). The GP is shown in Fig. 7. The following steps are taken:

1. A pipeline route is selected and a GP is plotted.
2. A series of potential pipe diameters is chosen with a range of sizes such that the average flow velocity of 6 ft/sec (1.83 m/sec) is bracketed for the design throughput of the pipe.
3. For each of these candidate pipes the slope of the HGL, $-h_f/L$, is computed. For the illustrative design problem we chose pipes of schedule 40 size with nominal diameters of 5, 6, 8, and 10 in. The results are shown in Table I.
4. The desired residual head at the terminal end of the pipeline is specified. This is governed by the requirements

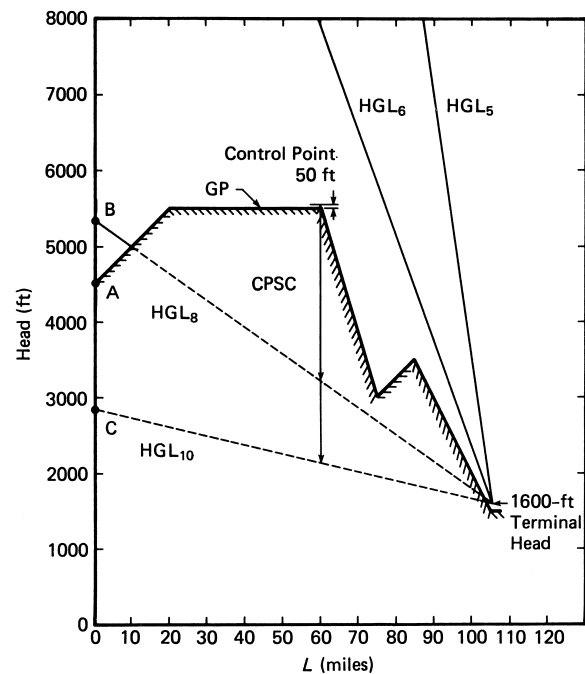


FIGURE 7 Ground profile (GP) plot showing initial hydraulic grade lines (HGLs) for pipes of different diameter. Eight- and 10-in. pipes (HGL₈, HGL₁₀) require additional control point static correction (CPSC) to clear the control point.

TABLE I Sample Design Problem Illustrating Hydraulic Grade Line Method

<i>D</i>		<i>V</i>		$-h_f/L$	
(in.)	(m)	(ft/sec)	(m/sec)	(ft/mile)	(m/km)
5	0.1270	9.41	2.87	338	64.0
6	0.1524	6.52	1.99	138	26.1
8	0.2032	3.76	1.15	35.6	6.74
10	0.2540	2.39	0.73	11.9	2.25

of the process to be fed by the pipeline system. For this case 100 ft is used.

5. Once the terminal end pressure head is decided on, it is used as an anchor point through which HGLs for the various pipes are drawn (lines of slope $-h_f/L$ passing through the terminal head point at the end of the line). This is illustrated for the candidate pipes in Fig. 7.

6. From the HGL/GP plot the control points are determined. These are points, such as mp-60 (mp refers to the mileage post along the horizontal axis) in Fig. 7, that must be cleared by the flatter HGLs in order to avoid slack flow conditions. These points, together with the slopes of the HGLs, determine the minimum heights to which the HGL must be raised at mp-0 and thus the pump head requirements for each pipe. Depending on the specific GP, there may be multiple control points.

7. The approximate number and size of pumps required for the job are estimated. This is done by determining the total hydraulic horsepower required for each pipe and dividing by a nominal pump head representative of pump types (of the order of 2000 psi for positive displacement

and 900 psi for centrifugal pumps). For the sample problem the results assuming 900-psi centrifugals are shown in Table II. In making the calculations in Table II a number of factors must be taken into account. The total Δp_f is the HGL $\Delta p/L$ times total length (105 miles). The CPSC is the control point static correction and represents the net head increase that must be added to the HGL at mp-0 to cause it to clear the GP at its critical interior control point by a minimum terrain clearance (taken here to be 50 ft). For the 8-in. pipe it is simply the vertical distance between the GP + 50 ft at the control point (mp-60) and the HGL at that point. This is so because at mp-0 the HGL starts at point *B* (see Fig. 7), which is above the GP. For the 10-in. pipe, however, the HGL actually starts at point *C*, which is below GP. Therefore, the CPSC is the vertical distance between GP + 50 ft and the HGL at mp-60 decreased by the negative head at mp-0 (point *C* minus point *A* in Fig. 7).

The significance of CPSC is that this is the additional head the pumps must produce in order to get the fluid up over the GP at the control point with a minimum terrain clearance. This, of course, results in the HGL terminating at mp-105 at a much higher head than the specified 1600-ft terminal end head. This excess head, also tabulated in Table II, must be wasted or “burned off” as friction. This can be accomplished in a number of ways, such as introducing an orifice plate, introducing a valve, or decreasing the pipe diameter. Depending on specific pipeline system conditions and economics, any of these alternatives may be desirable.

8. The hydraulic and actual horsepower required for the pumps are determined. The hydraulic horsepower (HHP) is given by

TABLE II Hydraulic Horsepower Calculations for Candidate Pipes

Nominal <i>D</i> (in.)	CP (miles) ^a	Total Δp_f (psi) ^b	CPSC (psi) ^{b,c}	Minimum pump pressure (psi) ^b	Approx. number of pump stations	HHP ^{d,e}	AHP ^f	AHP/PS ^g	Nominal HP/PS ^{g,h}	Actual head (ft) ⁱ	Excess head (ft) ^j
5	105	18,144	—	18,144	21	6211	8281	394	400	1714	—
6	105	7,408	—	7,408	9	2536	3381	376	400	1714	—
8	60	1,960	1196	3,156	4	1080	1440	360	400	1714	2348
10	60	639	895	1,534	2	525	700	350	350	767	3415

^a CP, control point.

^b 1 psi $\equiv$ 6894.8 Pa.

^c CPSC, control point static correction.

^d HHP (hydraulic horsepower) = Δp (psi) Q (gpm)/1714.

^e 1 hp $\equiv$ 745.7 W = 0.7457 kW.

^f AHP (actual horsepower) = HHP/Eff; Eff = 0.75 is assumed here.

^g PS, pump station.

^h Rounded up to nearest 50 hp.

ⁱ Based on nominal HHP/PS and 75% efficiency.

^j Head at mp-105 less terminal head for HGL, which clears interior CP by 50-ft minimum terrain clearance.

$$\text{HHP} = \Delta p_t(\text{psi})Q(\text{gpm})/1714, \quad (159)$$

while the actual horsepower (AHP) is HHP divided by the pump efficiency (here taken to be 0.75; actual values would be fixed by the vendor in a real case).

9. The nominal horsepower per pump station (HP/PS) is fixed. This is done by rounding the AHP/PS up to the next nearest 50 hp.

10. The actual head required is determined. This is done by taking the nominal HP/PS and computing the pump station pressure rise from Eq. (159).

11. The PS discharge head is determined. This is done by adding to the PS pressure rise just computed the net positive suction head (NPSH) of the pump as specified by the vendor. It is always wise to allow an additional head above this value as a safety factor. Here a 50-ft intake head has been assumed for illustrative purposes.

12. The PSs are located. Figure 8 contains the final results for the 8-in. pipe. The total PS discharge head is plotted above the GP at mp-0 (6264 ft in Fig. 8). From this point the HGL is plotted. When it reaches a point equal to the pump intake head (50 ft in this example) above the GP, the next PS is located (mp-20 in Fig. 8). Here the process repeated, and the PS pressure rise head is plotted above the HGL (7266 ft in Fig. 8). This process is repeated as many times as necessary to cause the HGL to clear

all control points and to terminate on the terminal mp. In this example no more PSs are required, and the HGL terminates at mp-105 at a head of 4240 ft. This is far too much head for the specified conditions of the design. The excess head (4240–1600 ft) must be consumed as friction, as already explained. In Fig. 8 the diameter is decreased to a 6-in. pipe at mp-79.2. This introduces the HGL for the 6-in. pipe, which now terminates at 1600 ft at mp-105 as desired.

13. The system is optimized. Steps 1–12 must be repeated for each candidate pipe. The entire set must then be cost optimized. For example, the design indicated by Fig. 8 will work hydraulically but is not optimum. We see that at mp-60, the interior control point, we have actually cleared GP by 342 ft. This is considerably more than the minimum 50-ft terrain clearance required and is therefore wasteful of pumping energy. The design can obviously be improved by a change in pump specifications and other details. This should be done for each candidate pipe. The final design to be selected is based on an economic minimum-cost evaluation.

The method just outlined and illustrated is route specific. It is very flexible and simple to use. It can also be easily computerized if the GP data can be fed in as numerical values. Here we have illustrated its use in the context of a cross-country pipeline, such as a crude oil, products, or perhaps slurry pipeline, which might be commonly encountered by chemical engineers. The method is completely adaptable to any hydraulic flow problem and could be used equally well for a short in-plant pumping system analysis. It can help the designer of flow systems to avoid sometimes subtle traps for slack flow and siphons that might not be immediately obvious if the mechanical energy equation is applied only once between the initial and final points of the flow system.

2. Pumps

Pumps come in a bewildering array of shapes, sizes, capacities, head characteristics, chemical and corrosion resistance features, materials of construction, and prime mover types. The choice of a specific pump for a specific application is best made in consultation with individual vendors who can provide detailed data about their product. Ultimately all choices are based on a cost optimization.

Pumps come basically in two types: (1) positive displacement and (2) centrifugal. As a rule of thumb, positive displacement pumps operate at high head but relatively low capacity. Centrifugals, on the other hand, operate at low head and high capacity. Typically, positive displacement (PD) pumps may operate at heads from 1 to 10,000 psi and from hundreds of gallons per minute to

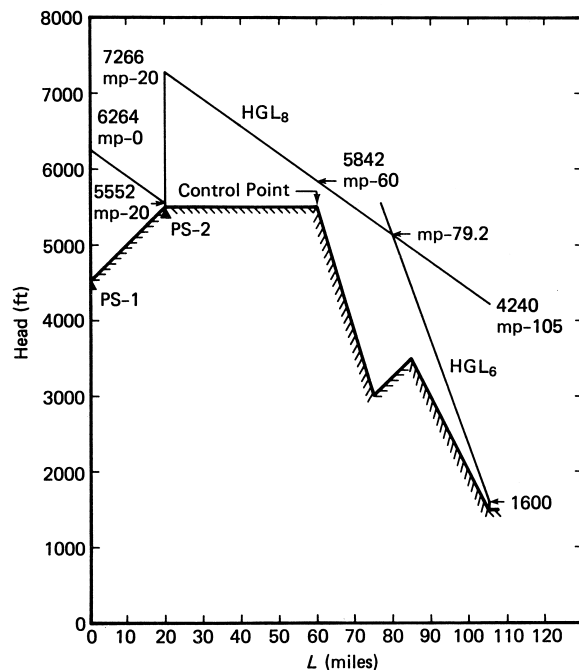


FIGURE 8 Hydraulic grade line (HGL) method design for pipe flow problem showing placement of pumping stations and change of diameter of pipe to handle excess head downstream of control point.

a fraction of a gallon per minute depending on the conditions. Typical centrifugal pumps may operate at heads of a few tens of feet to several hundreds of feet and capacities of several thousands of gallons per minute. It is possible to operate PD pumps in parallel or centrifugal pumps in series to achieve high head and high capacity. Some pump manufacturers also make “staged” centrifugal pumps, which are essentially multiple centrifugal pumps of identical head characteristics mounted on a common shaft and plumbed so as to permit the discharge of one to be the intake of the next stage.

a. Positive displacement pumps. Positive displacement pumps include gear pumps, piston pumps, plunger pumps, and progressing cavity pumps. All PD pumps have in common the fact that they are volumetric devices in which a fixed volume of fluid is drawn into the pump, pressurized, and discharged at high pressure into the line. As a result, the output is pulsatile, giving rise to a (sometimes violently) fluctuating discharge pressure. This necessitates the installation of pulsation dampeners at the discharge of all PD pumps in a large pumping installation to protect the system against heavy pressure surging.

Another feature of PD pumps is that, if the line for any reason becomes blocked, they simply continue forcing high-pressure fluid into the line and eventually break something if a precautionary rupture system has not been installed. Thus, a PD pump should be protected by a high-pressure shutoff sensor and alarm system and also a bypass line containing a rupture disk or pressure relief valve.

Figure 9 is a schematic illustration of a double-acting PD piston pump. The volumetric capacity of this device per stroke of the piston is given by

$$Q' = \left(\frac{1}{4}\pi D_p^2 L_s n - V_R\right) Ne, \quad (160)$$

where D_p is the diameter of the piston, L_s is the stroke length, $n = 1$ for a single-acting (only one side of the piston drives fluid on one-half of the stroke) or $n = 2$ for a double-acting (the piston drives fluid on both halves of the stroke)

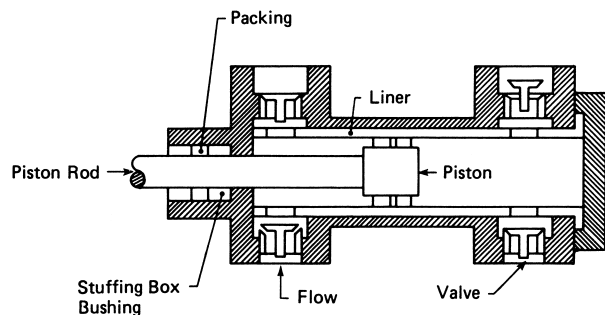


FIGURE 9 Schematic illustration of double-acting piston pump.

pump, V_R is the volume displaced by the rod in the double-acting case, N is the number of cylinders per pump, and e is a volumetric efficiency factor, usually 0.95–0.99.

The total volumetric capacity of the pump is

$$Q = \omega Q', \quad (161)$$

where ω is the frequency in strokes per time. As an illustration of the use of these equations, suppose that in the previous HGL sample design problem we had elected to use single-acting ($n = 1$, $V_R = 0$), triplex ($N = 3$) piston pumps with a 12-in. piston diameter and a 10-in. stroke. At a total throughput of 587 gpm we calculate $Q' = 3256$ in.³/stroke from Eq. (160) and from Eq. (161) we find $\omega = 41.6$ strokes per/minute. Armed with such information one can now seek a specific vendor. Adjustments in several of the design variables may need to be made to be compatible with vendor specifications.

A useful feature of the PD pump is that for a given power input Eqs. (159)–(161) allow the designer considerable flexibility in adjusting discharge pressure, cylinder capacity, and overall capacity. Positive displacement pumps are favorites on large-scale, high-pressure systems. Details of each of the various types of PD pump are best obtained from individual vendors.

b. Centrifugal pumps. The operation of centrifugal pumps is entirely different from that of PD pumps. The principle of operation involves spinning a circular vaned disk at high speed inside a casing. The resulting centrifugal force accelerates the fluid to high velocity at the tangential discharge port, where it stagnates against the fluid already in the pipe, creating high pressure as a result of Bernoulli's equation. As a result the discharge pressure of an ideal centrifugal pump is proportional to the square of the velocity of the impeller tip. In actual practice, however, frictional energy losses and turbulence within the pump result in a different relationship, which must be determined experimentally for each pump. This is routinely done by pump manufacturers, and the information is presented in the form of a pump head curve, such as that illustrated in Fig. 10.

Manufacturers' performance curves, such as those in Fig. 10, contain a great deal of useful information. Actual average head–capacity curves are shown for a number of impeller diameters. Also superimposed on these head curves are curves of constant efficiency. A third set of curves superimposed on the head curves are the NPSH requirement curves (dashed line in Fig. 10), which indicate the required NPSH at any given condition of operation. A fourth set of curves sometimes included are the BHP (brake horsepower) curves. BHP is the actual horsepower calculated in the previous HGL method illustration. It is the HHP divided by the efficiency.

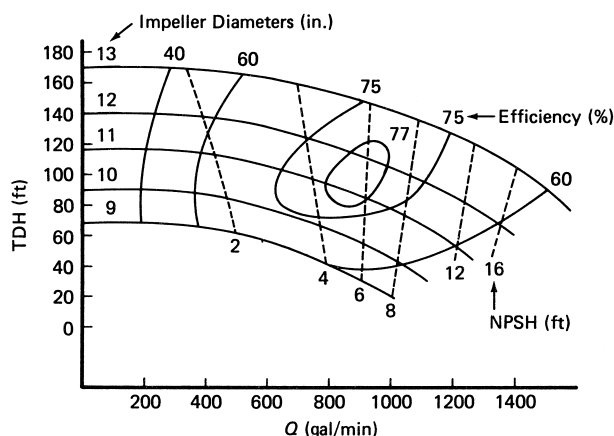


FIGURE 10 Typical centrifugal pump characteristic curves showing efficiency curves and NPSH (net positive suction head) for several impeller diameters.

We have discussed only a very small amount of information about pumps. A great deal more detail and practical operating information is available in books dealing with the selection of pumps. Space limitations preclude the inclusion of this detail here. In any specific application the user should consult with the pump vendors for assistance with details regarding materials of construction, installation, operation and maintenance, bearings, seals, valves, couplings, prime movers, and automatic controls.

3. Fitting Losses

From Eq. (63), the mechanical energy equation in head form, it is seen that, in the absence of a pump head, losses in a pipe system consist of pressure head changes, potential head changes, and velocity head changes. When fittings or changes in pipe geometry are encountered, additional losses occur.

It is customary to account for these losses either as pressure head changes over a length of pipe that produces the same frictional loss (hence an “equivalent length”) or in terms of a velocity head equivalent to the actual fitting head loss. In the earlier literature the equivalent length method was popular, with various constant equivalent lengths being tabulated for fittings of various types. More recently, however, it has been realized that flows through fittings may also be flow-rate dependent so that a single equivalent length is not adequate.

In the velocity head method of accounting for fitting losses, a multiplicative coefficient is found empirically by which the velocity head term $\langle v \rangle^2/2g$ is multiplied to obtain the fitting loss. This term is then added to the regular velocity head losses in Eq. (63). Extensive tables and charts of both equivalent lengths and loss coefficients and formulas for the effect of flow rate on loss coefficients

are published by manufacturers of fittings and valves. They are much too extensive to be reproduced here.

C. Noncircular Ducts

The mathematical analysis of flow in ducts of noncircular cross section is vastly more complex in laminar flow than for circular pipes and is impossible for turbulent flow. As a result, relatively little theoretical base has been developed for the flow of fluids in noncircular ducts. In order to deal with such flows practically, empirical methods have been developed.

The conventional method is to utilize the pipe flow relations with pipe diameter replaced by the hydraulic diameter,

$$D_H = 4A_c/P_w, \quad (162)$$

where A_c is the cross-sectional area of the noncircular flow channel and P_w is its wetted perimeter. For Newtonian flows this method produces approximately correct turbulent flow friction factors (although substantial systematic errors may result). It has not been tested for non-Newtonian turbulent flows. It can easily be shown theoretically to be invalid for laminar flow. However, for purposes of engineering estimating of turbulent flow one can obtain rough “ballpark” figures.

D. Drag Coefficients

When fluid flows around the outside of an object, an additional loss occurs separately from the frictional energy loss. This loss, called form drag, arises from Bernoulli’s effect pressure changes across the finite body and would occur even in the absence of viscosity. In the simple case of very slow or “creeping” flow around a sphere, it is possible to compute this form drag force theoretically. In all other cases of practical interest, however, this is essentially impossible because of the difficulty of the differential equations involved.

In practice, a loss coefficient, called a drag coefficient, is defined by the relation

$$F_D/A_c = C_D \rho v_\infty^2/2, \quad (163)$$

which is exactly analogous to the definition of f' , the Fanning friction factor. In this equation F_D is the total drag force acting on the body, A_c is the “projected” cross-sectional area of the body (a sphere projects as a circle, etc.) normal to the flow direction, ρ is the fluid density, v_∞ is the fluid velocity far removed from the body in the undisturbed fluid, and C_D is the drag coefficient.

In the case of Newtonian fluids, C_D is found to be a function of the particle Reynolds number,

$$\text{Re}_p = d_p v_\infty \rho / \mu, \quad (164)$$

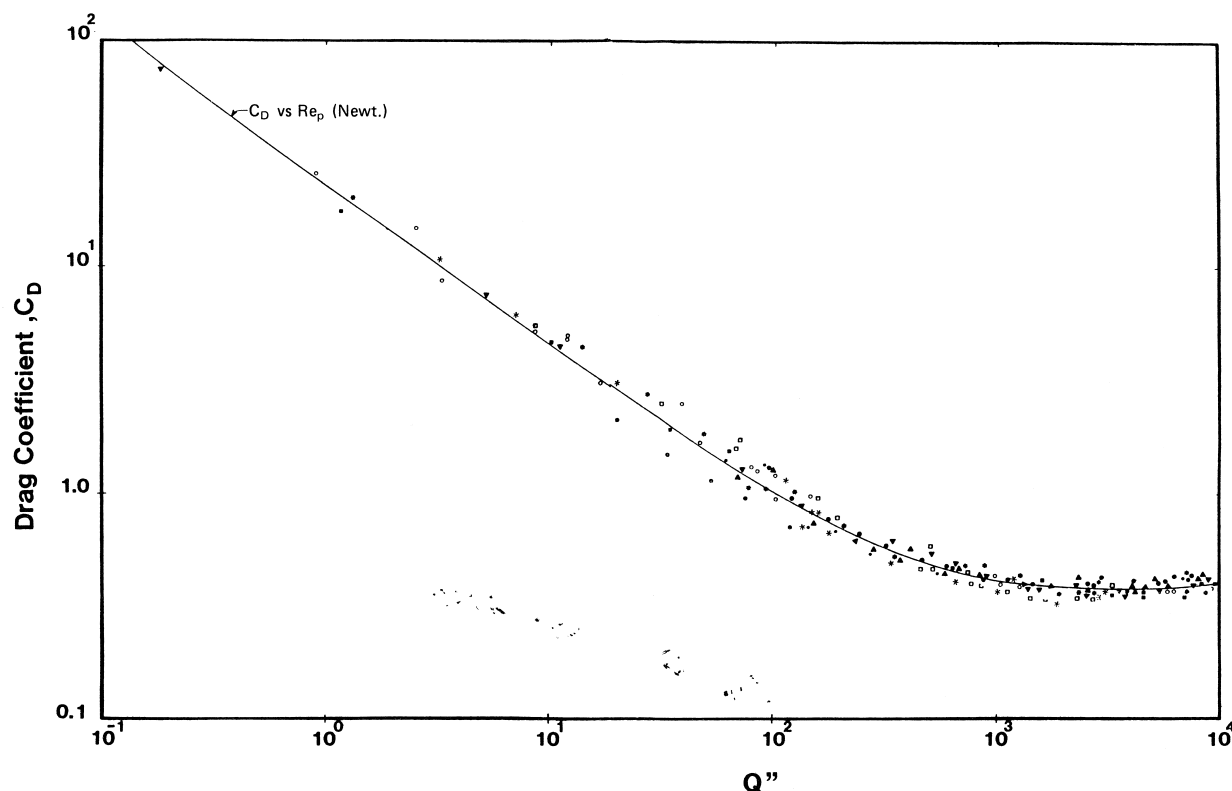


FIGURE 11 Generalized correlation of drag coefficient for Herschel–Bulkley model fluids; Q'' is defined by Eq. (165) and reduces to appropriate parameters for Bingham plastic, power law, and Newtonian fluid limits.

where d_p is the “effective” spherical diameter of the particle, v_∞ and ρ are as defined above, and μ is the viscosity of the fluid. The effective spherical diameter is the diameter of a sphere of equal volume. Also of importance are “shape” factors, which empirically account for the non-sphericity of real particles and for the much more complex flow distributions they engender.

Figure 11 is a plot of C_D as a function of a generalized parameter Q'' , defined by

$$Q'' = \frac{\text{Re}_{\text{pHB}}^2}{\text{Re}_{\text{pHB}} + (7\pi/24)\text{He}_{\text{pHB}}}, \quad (165)$$

where Re_{pHB} and He_{pHB} are the Reynolds number and Hedstrom number, respectively, for the Herschel–Bulkley rheological model defined as in the pipe flow case with D replaced by d_p .

This parameter is defined to accommodate Herschel–Bulkley model fluids. In the limit $\tau_0 = 0$, it reduces to an equivalent power law particle Reynolds number. In the limit $n = 1$, it reduces to a compound parameter involving the Bingham plastic particle Reynolds number and particle Hedstrom number. In both limits it reduces to the Newtonian particle Reynolds number. This correlation permits

one to determine drag coefficients for spheres in a wide variety of non-Newtonian fluids.

The curve in Fig. 11 has been represented by the following set of empirical equations to facilitate computerization of the iterative process of determining C_D ,

$$C_D = 24/Q'', \quad Q'' \leq 1 \quad (166)$$

$$C_D = \exp[q(\ln Q'')], \quad (167)$$

where the function $q(x)$ with $x = \ln(Q'')$ has the form

$$\begin{aligned} q(x) = & 3.178 - 0.7456x - 0.04684x^2 \\ & + 0.05455x^3 - 0.01796x^4 \\ & + 2.4619(10^{-3})x^{5*} - 1.1418(10^{-4})x^6. \end{aligned} \quad (168)$$

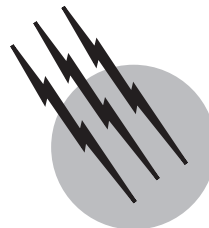
For $Q'' > 1000$, $C_D = 0.43$ is used. In the Newtonian limit, Eq. (166) is Stokes’ law.

SEE ALSO THE FOLLOWING ARTICLES

FLUID DYNAMICS • FLUID MIXING • LIQUIDS, STRUCTURE AND DYNAMICS • REACTORS IN PROCESS ENGINEERING • RHEOLOGY OF POLYMERIC LIQUIDS

BIBLIOGRAPHY

- Alexandrou, A. N. (2001). "Fundamentals of Fluid Dynamics," Prentice Hall, Englewood Cliffs, NJ.
- Batchelor, G. K. (2000). "An Introduction to Fluid Dynamics," Cambridge Univ. Press, Cambridge, U.K.
- Darby, R. (1996). "Chemical Engineering Fluid Mechanics," Dekker, New York.
- Dixon, S. L. (1998). "Fluid Mechanics and Thermodynamics of Turbomachinery," Butterworth-Heinemann, Stoneham, MA.
- Fuhs, A. E., ed. (1996). "Handbook of Fluid Dynamics and Fluid Machinery," 99E, Vols. 1–3, Wiley, New York.
- Garg, V. K. (1998). "Applied Computational Fluid Dynamics," Dekker, New York.
- Kleinstreuer, C. (1997). "Engineering Fluid Dynamics: An Interdisciplinary Systems Approach," Cambridge Univ. Press, Cambridge, U.K.
- Lin, C. A., Ecer, A., and Periaux, J., eds. (1999). "Parallel Computational Fluid Dynamics '98: Development and Applications of Parallel Technology," North-Holland, Amsterdam.
- Mc Ketta, J. J. (1992). "Piping Design Handbook," Dekker, New York.
- Middleman, S. (1997). "An Introduction to Fluid Dynamics: Principles of Analysis and Design," Wiley, New York.
- Sabersky, R. H., and Acosta, A. J. H. (1998). "Fluid Flow: A First Course in Fluid Mechanics," 4th ed., Prentice Hall, Englewood Cliffs, NJ.
- Siginer, D. A., De, D., and Kee, R. (1999). "Advances in the Flow and Rheology of Non-Newtonian Fluids," Elsevier, Amsterdam/New York.
- Sirignano, W. A. (1999). "Fluid Dynamics and Transport of Droplets and Sprays," Cambridge Univ. Press, Cambridge, U.K.
- Smits, A. J. (1999). "A Physical Introduction to Fluid Mechanics," Wiley, New York.
- Srivastava, R. C., and Leutloff, D. (1995). "Computational Fluid Dynamics: Selected Topics," Springer-Verlag, Berlin/New York.
- Upp, E. L. (1993). "Fluid Flow Measurements: Practical Guide to Accurate Flow Measurement," Gulf Pub., Houston.



Fluid Mixing

J. Y. Oldshue

Mixing Equipment Company, Inc.

- I. General Principles
- II. Scaleup Relationships
- III. Liquid–Solid Contacting
- IV. Gas–Liquid Contacting
- V. Liquid–Liquid Contacting
- VI. Blending
- VII. Fluid Motion
- VIII. Heat Transfer
- IX. Continuous Flow
- X. Pilot Plant Procedures
- XI. Computational Fluid Dynamics

GLOSSARY

Axial flow impellers Impellers that pump the fluid primarily in an axial direction when installed in a baffled mixing tank.

Chemical or mass transfer criteria Criteria for fluid mixing evaluation that involves measuring the rate of chemical reactions or rates of mass transfer across liquid, gas, or solid interfaces.

Computational fluid mixing Computer programs that use velocity data to calculate various types of flow patterns and various types of fluid mechanics variables used in analyzing a mixing vessel.

Fluidfoil impellers Axial flow impellers in which the blade shape and profile is patterned after airfoil concepts. The blade normally has camber and has a twist in toward the shaft with a rounded leading edge to pro-

duce a uniform velocity across the entire face width of the axial flow impeller.

Fluid shear rate Velocity gradient at any point in the mixing tank.

Fluid shear stress Product of shear rate and viscosity, which is responsible for many mixing phenomena in the tank.

Macroscale mixing, macroscale shear rates Particles on the order of 500–1000 μm and larger, or fluid elements of this size, respond primarily to average velocities at any point in the tank and are characterized as macroscale shear rate sensitive or related to macroscale mixing. Visual inspection of a tank normally yields information on the macroscale mixing performance.

Microscale shear rates, microscale mixing Any particles or fluid elements on the order of 100 μm or less respond primarily to the fluctuating velocity components

in turbulent flow or to shear rate elements on the order of that same size in viscous flow. Measurement of fluid mixing parameters at the microscale level involve the ability to resolve small elements of fluid parameters, as well as understanding the dissipation of energy at the microscale level.

Physical uniformity criteria Criteria for fluid mixing which involves physical sampling of tank contents or estimation of pumping of tank contents or estimation of pumping capacity and/or velocity values.

Radial flow impellers Impellers that pump fluid in essentially a radial direction when installed in a baffled mixing tank.

FLUID MIXING, as an engineering study, is the technology of blending fluid substances, including gases and solids, and is an integral process in most manufacturing operations involving fluid products. An important aspect of fluid mixing is the design and use of equipment. Fluids can be mixed in containers with rotating impellers or by means of jets, or in pipelines by internal baffles and passageways. Fluid mixing can involve primarily a physical suspension or dispersion that can be analyzed by the degree of composition or uniformity. Other operations may involve mass transfer across two-phase interfaces or chemical reactions in one or more phases. Information about microscale and macroscale mixing requirements are needed for process analysis and scaleup.

I. GENERAL PRINCIPLES

The power put into a fluid mixer produces pumping Q and a velocity head H . In fact all the power P which is proportional to QH appears as heat in the fluid and must be dissipated through the mechanism of viscous shear. The pumping capacity of the impeller has been measured for a wide variety of impellers. Correlations are available to predict, in a general way, the pumping capacity of the many impeller types in many types of configurations. The impeller pumping capacity is proportional to the impeller speed N and the cube of the impeller diameter D ,

$$Q \propto ND^3$$

The power drawn by an impeller in low- and medium-viscosity fluids is proportional to the cube of impeller speed N and the impeller diameter D to the fifth power,

$$P \propto N^3 D^5 \quad (1)$$

At higher viscosities other exponents are involved (discussed later).

If these three relations are combined it is seen that at constant power, one can vary the ratio of flow to impeller velocity head by a choice of D given by Eq. (2)

$$(Q/H)_P \propto D^{8/3} \quad (2)$$

This equation indicates that large-diameter impellers running at low speed give high flow and low shear rates, but small-diameter impellers running at high speed give us high shear rates and low pumping capacities. This important relationship also indicates that impeller velocity head is related in principle to macroscale shear rates. Thus, one has the ability to change the flow to fluid shear ratio.

In addition to the mathematical concepts brought out in Eq. (2), axial flow impellers, often applied as the pitched blade turbine (Fig. 1a), are inherently able to produce more flow at a given horsepower and impeller speed than radial flow impellers, typified by the flat blade disc turbine, shown in Fig. 1b. Some processes, such as blending and solids suspension, are affected primarily by pumping capacity and are not greatly influenced by the fluid shear rate. Therefore, it is typical in practice to use axial flow impellers when dealing with solids suspension and blending. Changes in D/T (where T is the tank diameter) can affect the flow-to-fluid-shear rate ratio relative to the various diameters:

$$(Q/H)_P \propto (D/T)^{8/3} \quad (3)$$

The introduction in recent years of the fluidfoil type of impeller, shown in Fig. 1c, further improves the pumping capacity of axial impellers and reduces the fluid shear rate by the actual design of the impeller blades themselves. Figure 2 illustrates the phenomena of the fluidfoil. The illustration indicates the desired flow pattern over the blade shape to minimize shear rate and maximize flow. For comparison, Fig. 2b shows fluid flow if the angle of the impeller blade in the fluid is not set at this optimum flow position. As shown in Fig. 2b, the turbulence and drag behind the impeller blade will cause increased power and reduced pumping efficiency. However, the turbulence and drag are not always a problem, because some processes require a certain level of turbulence and energy dissipation. In such processes, the use of the fluidfoil impeller type would not be as effective as other types that develop higher internal impeller zone shear rates.

There are now several varieties of fluidfoil impellers in use. The A310 is an effective impeller for the low viscosity region and has a negative response to viscosity at a Reynolds number of approximately 600. As shown in Fig. 3, the angle that the flow stream makes with the vertical starts to become greater than with the A200 impeller, so we can say effectively that the Reynolds number limitation on the A310 is approximately 200.



FIGURE 1 Three typical impellers for low and medium viscosity: (a) Axial flow, 45° blade (A200), (b) radial flow, disc turbine (R100), and (c) fluidfoil axial flow impeller (A300).

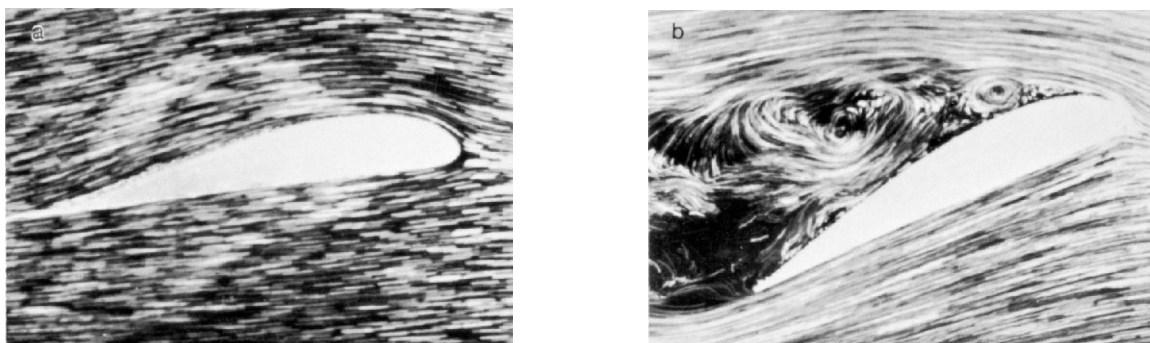


FIGURE 2 Typical air foil profiles. (a) Proper blade angle of attack for minimum drag and maximum flow for a given power. (b) Different blade angle of attack, giving higher drag coefficient and less flow per unit power.

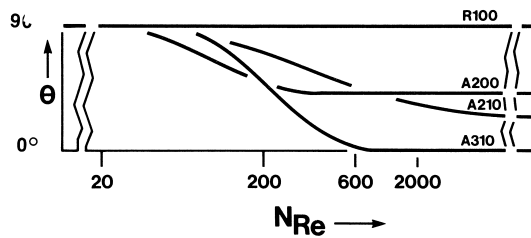


FIGURE 3 Changes in flow discharge angle with Reynold's number for four different impellers.

In order to carry this concept of fluidfoil impellers at a uniform velocity of discharge further, the A312 Impeller (Fig. 4) was developed and is used primarily in paper pulp suspensions. Carrying it further is the A320 Impeller (Fig. 5). The A320 has been studied particularly in the transitional area of traditional Reynolds numbers. This is shown in Fig. 6. This figure shows its performance and Reynolds numbers between 10 and 1,000.

For gas-liquid processes, the A315 impeller (Fig. 7) has been developed. This further increases the blade area and is used for gas-liquid applications.

The family of impellers shown here can be characterized by the solidity ratio, which is the ratio of area to blades to disc area circumscribing the impeller.

As shown in Fig. 8, the solidity ratio goes from 22% with the A310 up to 87% with the A315.

A. Shear Rate

There is a need to distinguish at this point how the shear rate in the impeller zone differs from the shear rate in the tank zone. To do this, however, one must carefully define shear rate and the corresponding concepts of macroscale shear rate and microscale shear rate. When one studies the localized fluid velocity through utilization of a small dimension probe, or as is currently used, a laser Doppler velocity meter device, one sees that at any point in the

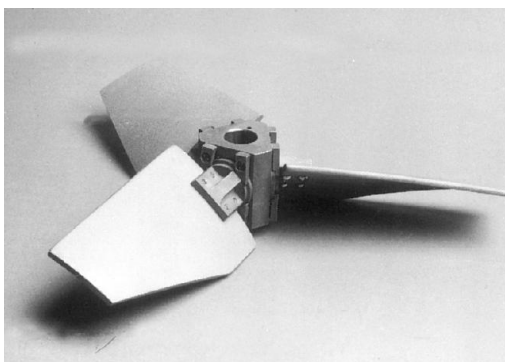


FIGURE 4 Photograph of A312 fluidfoil impeller.

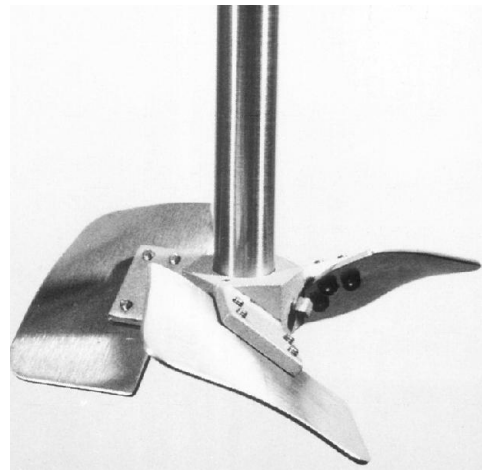


FIGURE 5 Photograph of A320 fluidfoil impeller.

stream of the tank there is a fluctuating velocity if we have turbulent flow (Fig. 9). From the curve in Fig. 9, one can calculate the average velocity at any point, as well as the fluctuating velocity above and below the average at this point. Figure 10 is a plot of the average velocity obtained from this curve. If these velocities are plotted at a constant discharge plane from the impeller, then the average impeller zone shear rate can be calculated. This average rate is really a macroscale shear rate, and it only refers to particles that have sizes much greater than $1000 \mu\text{m}$ that experience an effect from these shear rates. Also note that there is a maximum macroscale shear rate around the impeller. There are a variety of shear rates around the impeller, so that one needs to recognize the effect of each on a given process.

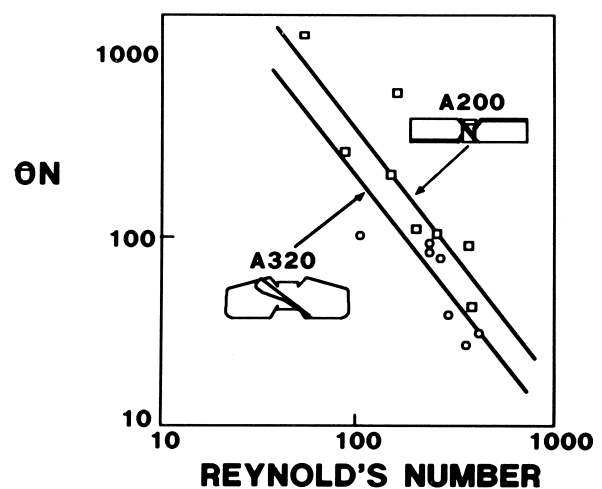


FIGURE 6 Effect of Reynolds number on blend number, θN , for the two impellers shown. θ , blend time; N , impeller rotational speed.

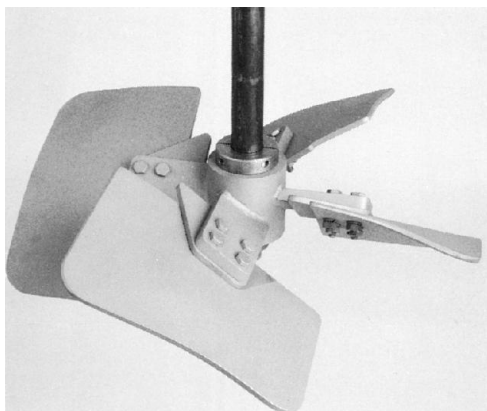


FIGURE 7 Photograph of A315 fluidfoil impeller.

In addition, the turbulent fluctuations set up a microscale type of shear rate. Microscale mixing tends to affect particles that are less than $100\ \mu\text{m}$ in size. The scaleup rules are quite different for macroscale controlled process in comparison to microscale. For example, in microscale processes, the major variables are the power per unit volume dissipated in various points in the vessel and the total average power per unit volume. In macroscale mixing, the energy level is important, as well as the geometry and design of the impeller blades and the way that they set up macroscale shear rates in the tank.

The fluidfoil impeller, shown in Fig. 1c, is often designed to have about the same total pumping capacity as the axial flow turbine (Fig. 1a). However, the flow patterns are somewhat different. The fluidfoil impeller has an axial discharge, while the axial flow turbine discharge tends to deviate from axial flow by $20\text{--}45^\circ$. Nevertheless at the same total pumping capacity in the tank, the tank shear rates are approximately equal. However, the axial flow fluidfoil turbine requires between 50 and 75% of the power required by the axial flow turbine. This results in a

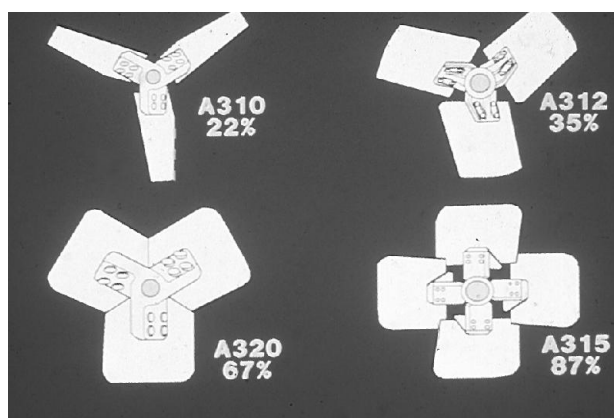


FIGURE 8 Solidty ratio of total blade area ratio to disc area of circumscribed circle at blade tips expressed as a percentage.

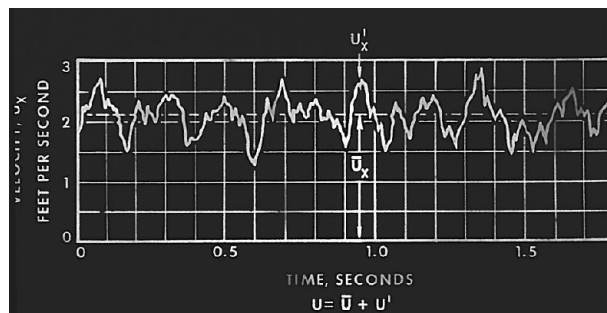


FIGURE 9 Schematic representation of turbulent flow recorded from a velocity probe as a function of time, showing average velocity and fluctuating velocity.

much smaller energy loss and dissipation in the impeller zone, and much lower microscale mixing in the impeller zone. There is also some difference in microscale mixing in the rest of the tank.

The lower horsepower is an important factor in the efficient design of axial flow or fluidfoil impellers. Such lower horsepower must be considered in the efficient design involving fluid velocity and overall macroscale mixing phenomena. On the other hand, if the process involves a certain amount of microscale mixing, or certain amounts of shear rate, then the fluidfoil impeller may not be the best choice.

Radial flow impellers have a much lower pumping capacity and a much higher macroscale shear rate. Therefore they consume more horsepower for blending or solids suspension requirements. However, when used for mass transfer types of processes, the additional interfacial area produced by these impellers becomes a very important factor in the performance of the overall process. Radial flow turbines are primarily used in gas-liquid, liquid-solid, or liquid-liquid mass transfer systems or any combinations of those.

B. Baffles and Impeller Position

Unbaffled tanks have a tendency to produce a vortex and swirl in the liquid. Such conditions may be wanted. Frequently, however, a good top-to-bottom turnover and the elimination of vortexing is needed. Therefore, baffles

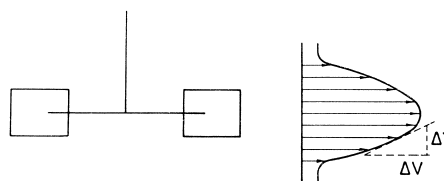


FIGURE 10 Illustration of average velocity from the radial discharge of a radial flow impeller, showing the definition of fluid shear rate ($\Delta V/\Delta Y$).

are used more often than not. Wall baffles for low-viscosity systems consist of four baffles, each $\frac{1}{12}$ of the tank diameter in width. Another method is to install an axial flow impeller type in an angular, off-center position, such that it gives good top-to-bottom turnover, avoids vortexing, and also avoids the use of baffles. Figure 11a shows a typical flow pattern for an unbaffled tank. A baffled tank axial radial flow is shown in Fig. 11b, and the angular off-center position is in Fig. 11c.

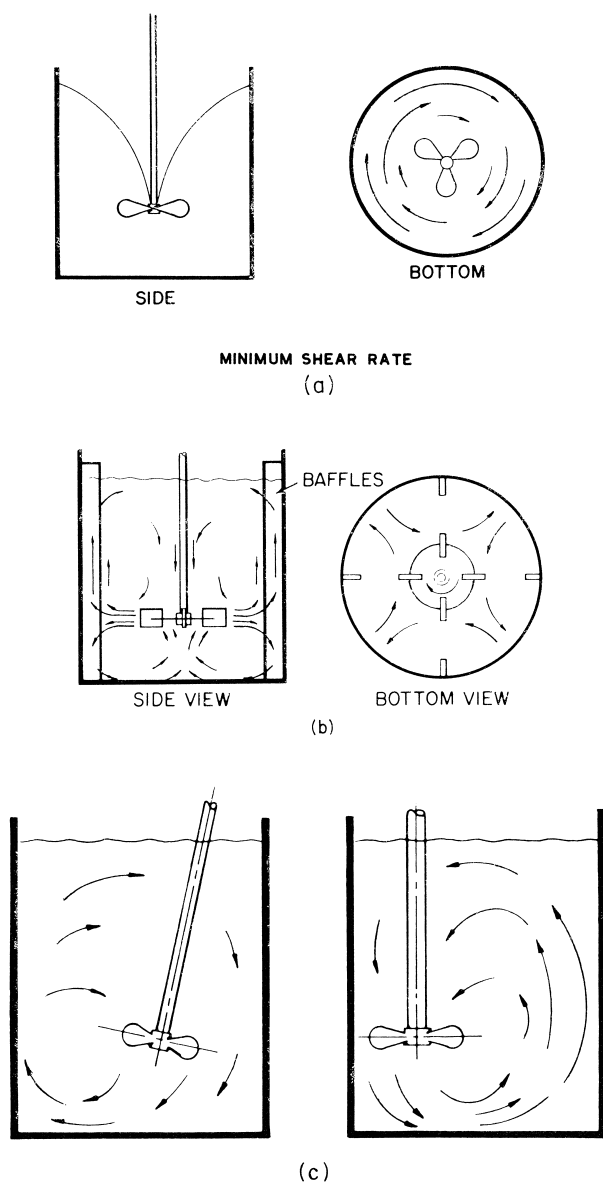


FIGURE 11 Effect of baffles in position on flow pattern. (a) Typical swirling and vortexing flow in a tank without baffles. (b) Typical top-to-bottom flow pattern with radial flow impellers with four wall baffles. (c) Typical angular off-center position for axial flow impellers to give top-to-bottom flow pattern to avoid swirl without the use of wall baffles.

TABLE I Elements of Mixer Design

Process design
Fluid mechanics of impellers
Fluid regime required by process
Scaleup; hydraulic similarity
Impeller power characteristics
Relate impeller hp, speed, and diameter
Mechanical design
Impellers
Shafts
Drive assembly

The need to use wall baffles to eliminate vortexing decreases as fluids become more viscous (5000–10,000 cP or more). But swirl will still be present if there are no baffles. Accordingly, quite often baffles of about one-half normal width are used in viscous materials. In such cases they are placed about halfway between the impeller and the wall.

C. Power Consumption

Table I shows the three areas of consideration in mixer design. The first area is process design, which will be covered in detail in succeeding pages. Process design entails determining the power and diameter of the impeller to achieve a satisfactory result. The speed is then calculated by referring to the Reynolds number–power number curve, shown in Fig. 12. Such a curve allows trial-and-error calculations of the speed once the fluid properties, P , D , and the impeller design are known.

D. Process Considerations

Table II gives a representation of the various types of mixing processes. The second column lists the nine basic areas of mixing: gas-liquid, liquid-solid, liquid-liquid, miscible liquid, fluid motion, and combinations of those. However, of more importance are the two adjacent columns. The first column includes physical processing, and has mixing criteria which indicate a certain degree of uniformity. The third column has chemical and mass transfer requirements, which involve the concept of turbulence, mass transfer, chemical reactions, and microscale mixing. Thus, there are summarized ten separate mixing technologies, each having its own application principles, scaleup rules, and general effect of process design considerations. In a complex process such as polymerization, there may possibly exist solids suspension, liquid-liquid dispersion, chemical reaction, blending, heat transfer, and other important steps. In general, it is more advantageous to break the process down into the component steps and consider the effect

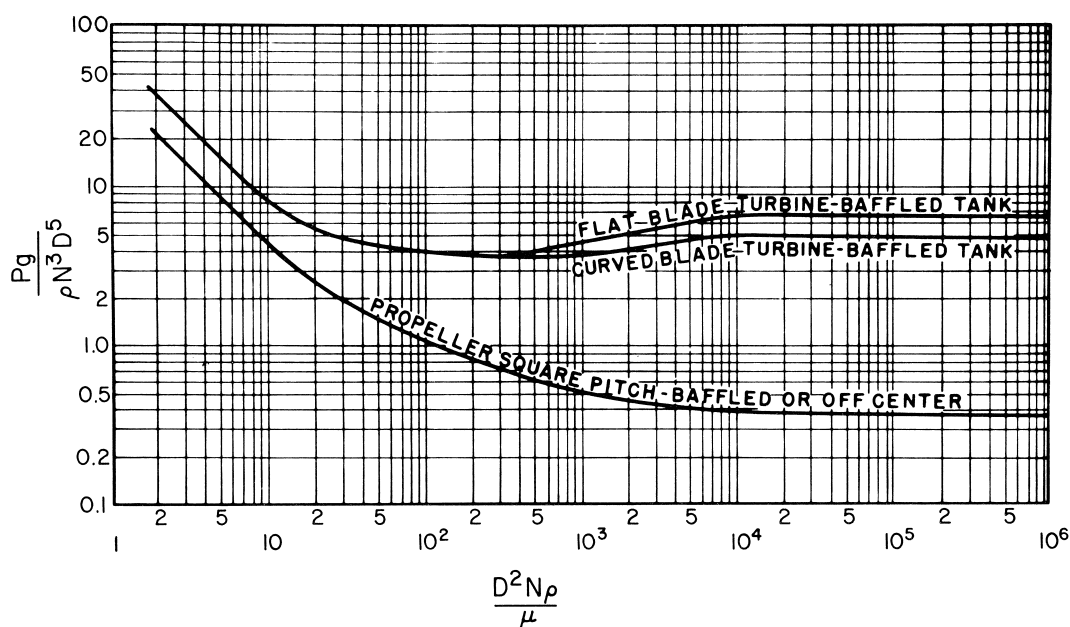


FIGURE 12 Reynolds number–power number curve for several impeller types: D , impeller diameter; N , impeller rotational speed; ρ , liquid density; μ , liquid viscosity; P , power; and g , gravity constant.

of mixing on each of these steps. One can then determine how the process will be affected by making changes in the mixer variables to the various mixing steps in the process. In scaleup, this is normally done by first determining the relative importance of the various steps, such as chemical reaction, mass transfer, blending, and so forth. The next step is to scaleup each of these steps separately to see the change on full-scale mixing. Later sections on scaleup and pilot planting will give some ideas on how scaleup affects typical performance variables.

Generally, heat transfer, blending, and solids suspension are governed primarily by the impeller's pumping capacity and not by fluid shear rates. Solid–liquid mass transfer, liquid–liquid mass transfer, and gas–liquid mass transfer have certain requirements for fluid shear in addi-

tion to pumping capacity: There are optimum ratios for those kinds of processes. There are many different combinations of impeller type and D/T ratios that can be used to get an optimum combination once the optimum flow to fluid shear is achieved. Thus, impeller design is not critical in terms of process performance but is critical in terms of economics of the overall mixer.

It is possible to use mixers as low head pumps by suitably installing them in a draft tube or above the orifice. They can then be used to pump large volumes of liquid at low heads.

The fluid mixing process involves three different areas of viscosity which affect flow patterns and scaleup, and two different scales within the fluid itself, macroscale and microscale. Design questions come up when looking at the design and performance of mixing processes in a given volume. Considerations must be given to proper impeller and tank geometry as well as the proper speed and power for the impeller. Similar considerations come up when it is desired to scaleup or scaledown and this involves another set of mixing considerations.

If the fluid discharge from an impeller is measured with a device that has a high frequency response, one can track the velocity of the fluid as a function of time (Fig. 9). The velocity at a given point in time can then be expressed as an average velocity ($\bar{v}$) plus fluctuating component (v'). Average velocities can be integrated across the discharge of the impeller and the pumping capacity normal to an arbitrary discharge plane can be calculated. This arbitrary discharge plane is often defined as the plane bounded by

TABLE II Characterization of Various Types of Mixing Processes

Physical processing	Application classes	Chemical process
Suspension	Liquid-Solid	Dissolving
Dispersion	Liquid-Gas	Absorption
Emulsions	Immiscible liquids	Extraction
Blending	Miscible liquids	Reactions
Pumping	Fluid motion	Heat transfers
	Liquid-solid-gas	
	Liquid-liquid-solid	
	Liquid-liquid-gas	
	Liquid-liquid-gas-solid	

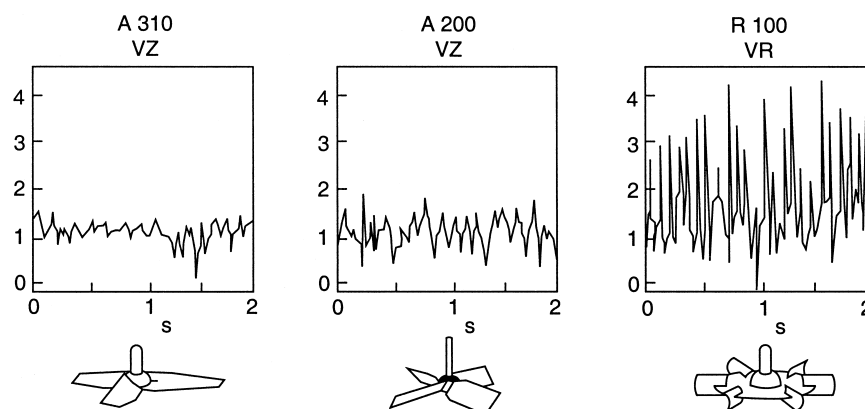


FIGURE 13 Velocity versus time for three different impellers.

the boundaries of the impeller blade diameter and height. Because there is no casing, however, an additional 10–20% of flow typically can be considered as the primary flow of an impeller.

The velocity gradients between the average velocities operate only on larger particles. Typically, these larger size particles are greater than $1000\ \mu\text{m}$. This is not a proven definition, but it does give a feel for the magnitudes involved. This defines macroscale mixing. In the turbulent region, these macroscale fluctuations can also arise from the finite number of impeller blades passing a finite number of impeller blades passing a finite number of baffles. These set up velocity fluctuations that can also operate on the macroscale.

Smaller particles primarily see only the fluctuating velocity component. When the particle size is much less than $100\ \mu\text{m}$, the turbulent properties of the fluid become important. This is the definition of the boundary size for microscale mixing.

All of the power applied by a mixer to a fluid through the impeller appears as heat. The conversion of power to heat is through viscous shear and is approximately 2500 Btu/hr/hp. Viscous shear is present in turbulent flow only at the microscale level. As a result, the power per unit volume is a major component of the phenomena of microscale mixing. At a $1\text{-}\mu\text{m}$ level, in fact, it doesn't matter what specific impeller design is used to apply the power.

Numerous experiments show that power per unit volume in the zone of the impeller (which is about 5% of the total tank volume) is about 100 times higher than the power per unit volume in the rest of the vessel. Making some reasonable assumptions about the fluid mechanics parameters, the root-mean-square (rms) velocity fluctuation in the zone of the impeller appears to be approximately 5–10 times higher than in the rest of the vessel. This conclusion has been verified by experimental measurements.

The ratio of the rms velocity fluctuation to the average velocity in the impeller zone is about 50% with many open impellers. If the rms velocity fluctuation is divided by the average velocity in the rest of the vessel, however, the ratio is on the order of 5–10%. This is also the level of rms velocity fluctuation to the mean velocity in pipeline flow. There are phenomena in microscale mixing that can occur in mixing tanks that do not occur in pipeline reactors. Whether this is good or bad depends upon the process requirements.

Figure 13 shows velocity versus time for three different impellers. The differences between the impellers are quite significant and can be important for mixing processes.

All three impellers are calculated for the same impeller flow, Q , and same diameter. The A310 (Fig. 2) draws the least power, and has the least velocity fluctuations. This gives the lowest microscale turbulence and shear rate.

1. The A200 (Fig. 3) shows increased velocity fluctuations and draws more power.
2. The R100 (Fig. 4) draws the most power and has the highest microscale shear rate.
3. The proper impeller should be used for each individual process requirement.

The velocity spectra in the axial direction for an axial flow impeller A200 is shown in Fig. 14. A decibel correlation has been used in Fig. 5 because of its well-known applicability in mathematical modeling as well as the practicality of putting many orders of magnitude of data on a reasonably sized chart. Other spectra of importance are the power spectra (the square of the velocity) and the Reynolds stress (the product of the R and Z velocity components), which is a measure of the momentum at a point.

The ultimate question is this: How do all of these phenomena apply to process design in mixing vessels? No one

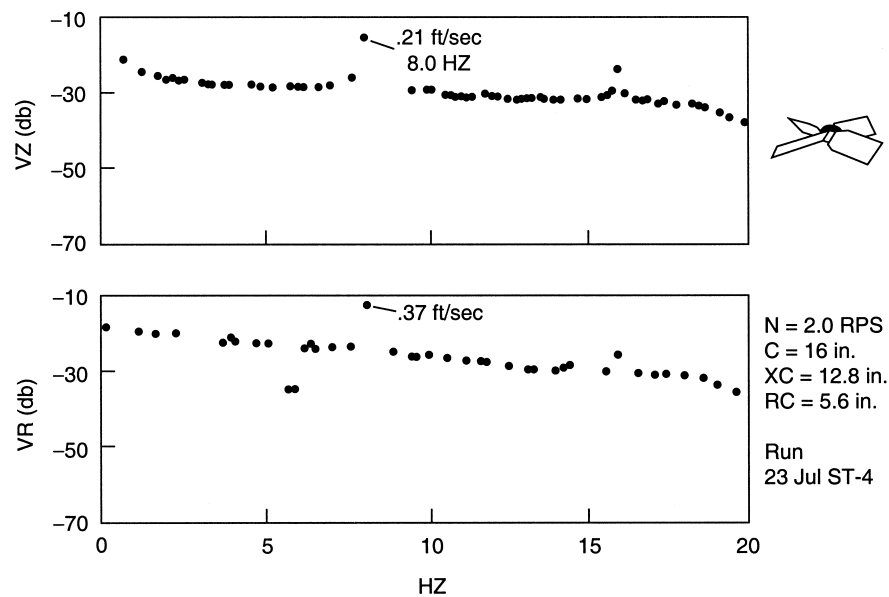


FIGURE 14 The velocity spectra in the axial direction for an axial impeller A200.

today is specifying mixers for industrial processes based on meeting criteria of this type. This is largely because processes are so complex that it is not possible to define the process requirements in terms of these fluid mechanics parameters. If the process results could be defined in terms of these parameters, sufficient information probably exists to permit the calculation of an approximate mixer design. It is important to continue studying fluid mechanics parameters in both mixing and pipeline reactors to establish what is required by different processes in fundamental terms.

Recently, one of the most practical results of these studies has been the ability to design pilot plant experiments (and, in many cases, plant-scale experiments) that can establish the sensitivity of process to macroscale mixing variables (as a function of power, pumping capacity, impeller diameter, impeller tip speeds, and macroscale shear rates) in contrast to microscale mixing variables (which are relative to power per unit volume, rms velocity fluctuations, and some estimation of the size of the microscale eddies).

Another useful and interesting concept is the size of the eddies, L , at which the power of an impeller is eventually dissipated. This concept utilizes the principles of isotropic turbulence developed by Komolgoroff [1]. The calculations assume some reasonable approach to the degree of isotropic turbulence, and the estimates do give some idea as to how far down in the microscale size the power per unit volume can effectively reach

$$L = (\nu^3/e)^{1/4}$$

where ν is the dynamic viscosity and e is the power per unit volume.

II. SCALEUP RELATIONSHIPS

Scaleup involves determining the controlling factors in a process, the role that mixing plays, and the application of a suitable scaleup technique. In this section, the general scaleup relationships will be presented, and the particular types of processes involved will be covered. Section X will cover pilot planting, how runs are made to determine the controlling factor, and how to choose a suitable design relationship for that situation.

Table III is a key for understanding scaleup relationships. In the first column are listed many design variables involved in mixing processes. These include power, power per unit volume, speed, impeller diameter, impeller

TABLE III Properties of a Fluid Mixer on Scaleup

Property	Pilot scale		Plant scale (17.280 liters)		
	(80 Liters)				
P	1.0	216	7776	36	0.16
P/Vol	1.0	1.0	36	0.16	.0007
N	1.0	0.3	1.0	0.16	.03
D	1.0	6.0	6.0	6.0	6.0
Q	1.0	65	216	36	6.0
Q/Vol	1.0	0.3	1.0	0.16	.03
ND	1.0	1.8	6.0	1.0	0.16
$\frac{ND^2\rho}{\mu}$	1.0	10.8	36	5.8	1.0

pumping capacity, pumping capacity per unit volume, impeller tip speed, and Reynolds number. In the second column, all these values are given a common value of 1.0, to examine the changes relative to each other on scaleup. In the remaining columns, a specific variable is held constant. When power per unit volume is constant, the speed drops, the flow increases, but the flow per unit volume decreases. The impeller tip speed goes up, and the Reynolds number goes up. It is quite apparent that the ratio of all the variables cannot be maintained as in the pilot plant. In addition, it appears that the maximum impeller zone shear rate will increase, while the circulating time and the impeller Reynolds number increase. This means that the big tank will be much different from the small tank in several potentially key parameters. When the flow per unit volume is held constant, the power per unit volume increases in proportion to the square of the tank diameter ratio. This is possible to do but is normally impractical.

When the impeller tip speed is held constant, the same maximum shear rate is maintained. However, the average impeller shear rate related to impeller speed drops dramatically, and the power per unit volume drops inversely to the tank size ratio. In general, this is a very unconservative scaleup technique and can lead to insufficient process results on full scale.

The final column shows results for a constant Reynolds number, which requires that the total power decrease on scaleup. This is not normally practical, and therefore we must accept an increased Reynolds number on scaleup. To complete this picture refer to Fig. 15, which shows that the maximum impeller zone shear rate increases, while the average impeller zone shear rate decreases during scaleup.

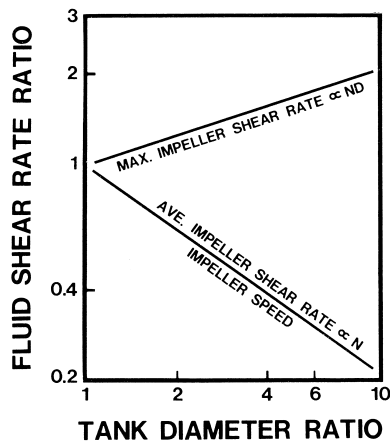


FIGURE 15 Schematic illustration of the increase in maximum impeller zone macroscale shear rate and a decrease of average impeller zone macroscale shear rate as tank size is increased, illustrating a wider distribution of shear rates in a large tank than in a small tank. The figure is based on a constant power/volume ratio and geometric similarity between the two tanks.

Both Table III and Fig. 15 are based on geometric similarity. One way to modify these marked changes in mixing parameters and scaleup is to use nongeometric similarity on scaleup. The problem is that the big tank has a much longer blend time than the small tank. The large tank has a greater variety of shear rates and has a higher Reynolds number than a small tank. These effects can be greatly modified by using nongeometric impellers in the pilot plant.

A. Role of Dynamic and Geometric Similarity

Equations (4) and (5) show the relationship for geometric and dynamic similarity, respectively, and illustrates four basic fluid force ratios.

$$\frac{X_M}{X_P} = X_R \quad (3)$$

$$\frac{(F_I)_M}{(F_I)_P} = \frac{(F_\mu)_M}{(F_\mu)_P} = \frac{(F_g)_M}{(F_g)_P} = \frac{(F_\sigma)_M}{(F_\sigma)_P} = F_R \quad (4)$$

where F is the fluid force, I the inertia force, μ , the viscous force, g the gravitational force, σ the surface tension force, M the model, P the prototype, and R the ratio. Subscript I is the inertia force added by the mixer, and it is desirable that it remain constant between the model M and the prototype P . Three fluid forces oppose the successful completion this process: viscosity, gravity, and fluid interfacial surface tension. It is impossible to keep these force ratios constant in scaleup with the same fluid. Therefore, we must choose two to work with. This, then, has led to the concept of dimensionless numbers, shown below.

$$\frac{F_I}{F_v} = N_{Re} = \frac{ND^2\rho}{\mu}$$

$$\frac{F_I}{F_g} = N_{Fr} = \frac{N^2D}{g}$$

$$\frac{F_I}{F_\sigma} = N_{We} = \frac{N^2D^3\rho}{\sigma}$$

in which the Reynolds number (the ratio of inertia force to viscous force) is shown, as well as the Froude number and the Weber number. The Reynolds number and power number curve have been discussed, in which the power number is the ratio of inertia force to acceleration. To illustrate the characteristics of dimensionless numbers in mixer scaleup, examine the case of blending. We can express blending performance in terms of blend time multiplied by impeller speed, which gives a dimensionless process group. This is shown in Fig. 16 and gives a good correlation against the Reynolds number. However, for the other thousands of applications that are designed each year, there is normally no good way to write a

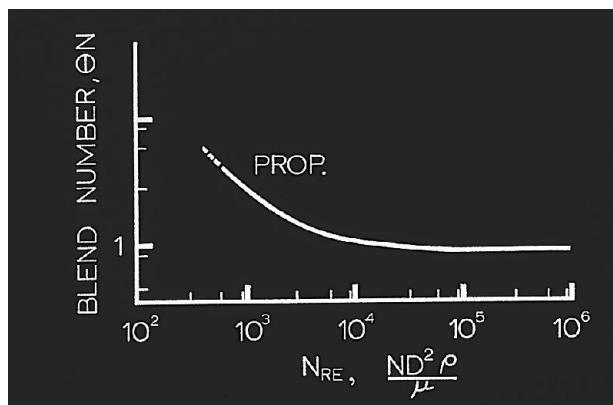


FIGURE 16 Typical dimensionless process correlation of blend number θN versus Reynolds number.

dimensionless group around the process result. For example, including polymerization yield, including productivity of a fermentation process, or incorporating the rate of absorption of flue gas into caustic does not allow a dimensionless type of process grouping. Thus, it is not practical to deal with dimensionless numbers when we do not have the ability to write a dimensionless group around the mixing process result.

There are as many potential scaleup parameters as there are individual process mixing results. However, we can make some generalizations which are very helpful in dealing with actual mixing problems, but for reliable scaleup, some experimental verification of the scaleup method to be used is desirable.

For example, it is found that the mass transfer coefficient, $K_G a$, for gas-liquid processes, is mostly a function of the linear superficial gas velocity and the power per unit volume with the constant D/T ratio for various size tanks. This is because the integrated volumetric mass transfer coefficient over the entire tank can be quite similar in large and small tanks even though the individual bubble size, interfacial area, and mass transfer coefficient can vary at specific points within the small and large tanks.

It has also been observed that suspension and blending of slurries operating in the hindered settling range (such as with particle sizes on the order of 100 mesh or smaller) tend to show a decreasing power per unit volume on scaleup. When this relationship is used, the blend time for the large tank is much longer than it is for the small tank. Blend time is not a major factor in a large slurry holding tank in the minerals processing industry, and therefore, that factor is not an important one to maintain on scaleup.

For homogeneous chemical reactions, most of the effect of the mixer occurs at the microscale level. Microscale mixing is largely a function of the power per unit volume, and maintaining equal power per unit volume gives similar

chemical reaction requirements for both small and large tanks.

Some processes are governed by the maximum impeller zone shear rate. For example, the dispersion of a pigment in a paint depends upon the maximum impeller zone shear rate for the ultimate minimum particle size. However, when constant tip speed is used to maintain this, the other geometric variables must be changed to maintain a reasonable blend time, even though process results on full scale will probably take much longer than those on small scale.

Two aspects of scaleup frequently arise. One is building a model based on pilot plant studies that develop an understanding of the process variables for an existing full-scale mixing installation. The other is taking a new process and studying it in the pilot plant in such a way that pertinent scaleup variables are worked out for a new mixing installation.

There are a few principles of scaleup that can indicate what approach to take in either case. Using geometric similarity, the macroscale variables can be summarized as follows:

- Blend and circulation times in the large tank will be much longer than in the small tank.
- Maximum impeller zone shear rate will be higher in the larger tank, but the average impeller zone shear rate will be lower; therefore, there will be a much greater variation in shear rates in a full-scale tank than in a pilot unit.
- Reynolds numbers in the large tank will be higher, typically on the order of 5–25 times higher than those in a small tank.
- Large tanks tend to develop a recirculation pattern from the impeller through the tank back to the impeller. This results in a behavior similar to that for a number of tanks in a series. The net result is that the mean circulation time is increased over what would be predicted from the impeller pumping capacity. This also increases the standard deviation of the circulation times around the mean.
- Heat transfer is normally much more demanding on a large scale. The introduction of helical coils, vertical tubes, or other heat transfer devices causes an increased tendency for areas of low recirculation to exist.
- In gas-liquid systems, the tendency for an increase in the gas superficial velocity upon scaleup can further increase the overall circulation time.

What about the microscale phenomena? These are dependent primarily on the energy dissipation per unit volume, although they must also be concerned about the

energy spectra. In general, the energy dissipation per unit volume around the impeller is approximately 100 times higher than in the rest of the tank. This results in an rms velocity fluctuation ratio to the average velocity on the order of 10:1 between the impeller zone and the rest of the tank.

Because there are thousands of specific processes each year that involve mixing, there will be at least hundreds of different situations requiring a somewhat different pilot plant approach. Unfortunately, no set of rules states how to carry out studies for any specific program, but here are a few guidelines that can help one carry out a pilot plant program.

- For any given process, take a qualitative look at the possible role of fluid shear stresses. Try to consider pathways related to fluid shear stress that may affect the process. If there are none, then this extremely complex phenomena can be dismissed and the process design can be based on such things as uniformity, circulation time, blend time, or velocity specifications. This is often the case in the blending of miscible fluids and the suspension of solids.
- If fluid shear stresses are likely to be involved in obtaining a process result, then one must qualitatively look at the scale at which the shear stresses influence the result. If the particles, bubbles, droplets, or fluid clumps are on the order of 1000 μm or larger, the variables are macroscale and average velocities at a point are the predominant variable.

When macroscale variables are involved, every geometric design variable can affect the role of shear stresses. They can include such items as power, impeller speed, impeller diameter, impeller blade shape, impeller blade width or height, thickness of the material used to make the impeller, number of blades, impeller location, baffle location, and number of impellers.

Microscale variables are involved when the particles, droplets, baffles, or fluid clumps are on the order of 100 μm or less. In this case, the critical parameters usually are power per unit volume, distribution of power per unit volume between the impeller and the rest of the tank, rms velocity fluctuation, energy spectra, dissipation length, the smallest microscale eddy size for the particular power level, and viscosity of the fluid.

- The overall circulating pattern, including the circulation time and the deviation of the circulation times, can never be neglected. No matter what else a mixer does, it must be able to circulate fluid throughout an entire vessel appropriately. If it cannot, then that mixer is not suited for the tank being considered.

Qualitative and, hopefully, quantitative estimates of how the process result will be measured must be made in advance. The evaluations must allow one to establish the importance of the different steps in a process, such as gas-liquid mass transfer, chemical reaction rate, or heat transfer.

- It is seldom possible, either economically or time-wise, to study every potential mixing variable or to compare the performance of many impeller types. In many cases, a process needs a specific fluid regime that is relatively independent of the impeller type used to generate it. Because different impellers may require different geometries to achieve an optimum process combination, a random choice of only one diameter of each of two or more impeller types may not tell what is appropriate for the fluid regime ultimately required.
- Often, a pilot plant will operate in the viscous region while the commercial unit will operate in the transition region, or alternatively, the pilot plant may be in the transition region and the commercial unit in the turbulent region. Some experience is required to estimate the difference in performance to be expected upon scaleup.
- In general, it is not necessary to model Z/T ratios between pilot and commercial units, where Z is the liquid level.
- In order to make the pilot unit more like a commercial unit in macroscale characteristics, the pilot unit impeller must be designed to lengthen the blend time and to increase the low maximum impeller zone shear rate. This will result in a greater range of shear rates than is normally found in a pilot unit.

All of these conditions can be met using smaller D/T ratios and narrower blade heights than are used normally in a pilot unit. If one uses the same impeller type in both the pilot and commercial units, however, it may not be possible to come close to the long blend time that will be obtained in the commercial unit. Radial flow impellers can be excellent models in a pilot plant unit for axial flow impellers in a commercial unit.

III. LIQUID–SOLID CONTACTING

Solids suspension involves producing the required distribution of solids in the tank and is essentially a physical phenomenon. The criterion is normally a physical description of the degree of uniformity required in the suspension. A key variable for solids suspension is the settling velocity of the solids. This is usually measured by timing the fall velocity of individual solid particles in a defined depth of

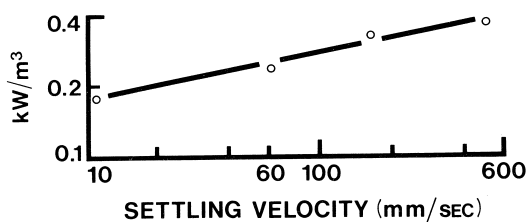


FIGURE 17 Effect of settling velocity to achieve a 60% suspension of particle sizes when there is a mixture of particle sizes.

mother liquor. When there is a wide range of particle sizes, there may well be a wide range of settling velocities.

Much of the literature is based on experimental data with similarly sized particles and observations of the speed required to keep particles in motion with at most 1 or 2 sec of rest on the bottom of the tank. This is done by visually observing solids in a transparent tank. This, of-course, means that relatively small-scale experiments are conducted and that this particular criterion cannot be used for studies in large-sized tanks or in field tests.

Sizing procedures to design a mixer for one closely sized particle settling velocity are modified considerably when there are other solids present. Figure 17 shows the effect of settling velocity on power when there are other solids present in the system. The slope is much less pronounced than it is when a single particle size alone is being suspended.

Much of the literature correlations for solids suspension are based on the so-called critical impeller speed. Attempts to duplicate experiments between various investigators often yield deviations of ± 30 –50% from the critical speed shown by other investigators. Because power is proportional to speed cubed, power varies on the order of 2 to 3 times, which is not sufficiently accurate for industrial full-scale design. Therefore, many approximate, conservative estimates have been made in the literature as general guidelines for choosing mixers for solids suspension. Table IV is one such guideline for solid particles of a closely sized nature.

The study of solids suspension in quantitative terms normally involves a method of sampling. Typically, samples are withdrawn from the side of the mixing tank through openings or tubes inserted into the vessel wall. It may also be done by submerging a container and quickly removing

the top and replacing it, allowing the slurry to flow into the container. Neither of these methods gives the absolute percentage of solids at the measurement point. In the case of a tube, the withdrawal velocity of the tube can affect the percent of solids that comes out of the discharged slurry as well as the orientation of the tube relative to the flow pattern in the tank. In the case of the sample container, its location, fill rate, and other variables can affect the actual solids composition measured compared to the actual. On the other hand, as long as measurement techniques are consistent, a reliable effect of mixer variables can be determined, which is of value in predicting operating conditions for full-scale units. One such test on a pilot plant scale yields data shown in Fig. 18, which shows the difference in axial and radial flow of solids suspension characteristics and indicates, as we mentioned previously, that axial flow impellers require less horsepower for the same degree of solids suspension.

The use of the new type of fluidfoil impeller has reduced the power required for solids suspension to about one-half to two-thirds of the values formerly used with 45° pitch blade turbines.

In continuous flow, the only point in the tank that must be equal to the feed composition for steady-state operation is the drawoff point. Thus, if the drawoff point is at the bottom, middle, or top of the tank, different average tank compositions can result, even though the composition of the entrance and exit streams are the same. If the mixer is large enough to provide complete uniformity of all the solids, including the coarse particles as well as the fine particles, then the drawoff point does not make any difference in the composition of the tank. However, if the mixer is designed only to just suspend the solids to the drawoff point, then tank compositions vary widely, depending upon the drawoff conditions.

Many times a fillet can be left in a tank, which will reduce the horsepower considerably for what will be required to completely clean out the last corners of a flat bottom tank. Depending upon the value of the solids in the process, they may either be left to form their own fillet, or the tank may be streamlined by using concrete or other materials to give a more streamlined shape.

When solids increase in percentage, the effect is to make the process requirement more difficult, and a curve similar to that in Fig. 19 results, until a point which often occurs around 40–50% by weight solids, at which there may be a discontinuity. At this point, the viscosity of the slurry is becoming a parameter, which reduces the settling velocity and, thus, minimizes its importance as a criterion to one in which we are essentially blending and providing motion through a pseudo-plastic fluid. Then as the solids percentage gets up toward 70 or 80% (and this point can be normalized by relating it to the percentage of the ultimate

TABLE IV Motor Horsepower for Estimating Purposes for Solids Suspension^a

Settling velocity	1'/min	2'/min	4'/min
Off bottom	1	2	5
Uniform	1.5	5	15

^a 15,000 gal tank; $D/T = 0.33$. $C/D = \frac{1}{2}$; axial flow turbine; 1–20% solids by weight.

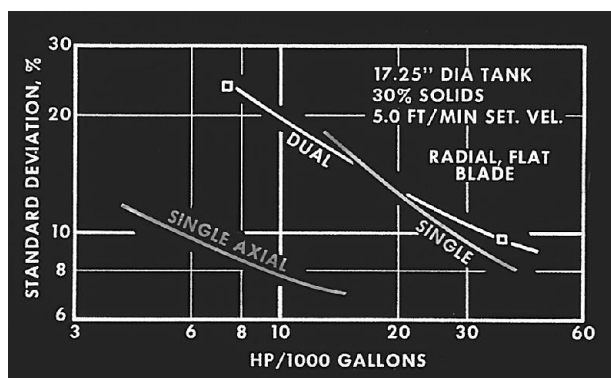


FIGURE 18 Typical comparison of power required for axial flow impeller compared to radial flow impellers in solids suspension.

settled solids), power becomes extremely high and approaches infinity where there is no supernatant liquid left in the tank. To evaluate this effect, a mixing viscosimeter is valuable, in which the slurry is agitated at the same time that the viscosity is measured, so that the measurement gives a reasonable value for the overall slurry.

A. Typical Mass Transfer Processes

Many processes involve criteria other than solids suspension, for example, crystallization, precipitation, and many types of leaching and chemical reactions. In crystallization, the shear rate around the impeller and other mixing variables can affect the rate of nucleation, and can affect the ultimate particle size. In some cases, the shear rate can be such that it can break down forces within the solid particle and can affect the ultimate particle size and shape. There are some very fragile precipitate crystals that are very much affected by the mixer variables.

In leaching, there usually is a very rapid leach rate that occurs when the mineral is on the surface of the particle, but many times the internal diffusion of the solid through the solid particle becomes controlling, and mixer variables do not affect the leaching rate beyond that point. In studying the effect of mixing on leaching processes, it is normally desirable to run separate experiments with

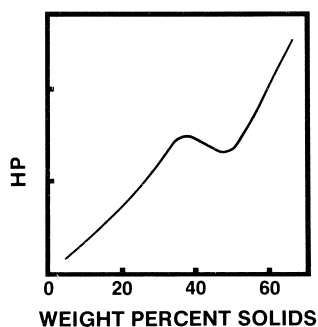


FIGURE 19 Increase of process horsepower versus weight percent solids, showing discontinuity when criteria changes from solids suspension to pseudo-plastic blending.

the fine particle sizes, the average particle sizes, and the coarse particle sizes, since the leaching curves are often quite different. In addition, the suspension of the fine, average, and coarse particles may be different in the leach tank due to the fact that all these particles may not be completely uniform throughout the system.

Typical design of an industrial leaching system looks at the extraction rate versus time and power level and determines the optimum combination of tank size and mixer horsepower in terms of return on the product leached.

B. Industrial Examples

One area for industrial studies is the whole area of slurry pipelines. Coal is by far the most common material in slurry pipelines, but other pipelines, but other pipelines include iron ore and potash. In large volume solid suspension applications, there is a considerable trade-off between volume of a tank, mixer horsepower, shape of a tank, and many other areas of cost consideration that are important in overall design. In addition to the tanks in these sorts of slurry systems, it must be capable of incorporating slurries into water or vice versa to either increase or decrease the solids concentration of a given system.

Another industrial application is mixing of paper pulp and slurries. An entire technology exists for this fluid, which is quite unique compared to other liquid-solid systems. Basically, there is a question of whether to use baffles, comparison of both top-entering and side-entering mixers, as well as the very large effect of type of paper pulp and the consistency of the paper pulp in the vessel. Other examples include fermentation, in which there is a biological solid producing the desired product, and the role of fluid shear rates on the biological solids is a critical consideration as well as the gas-liquid mass transfer (see Section VI).

Another class of applications is the high shear mixers used to break up agglomerates of particles as well as to cause rapid dissolving of solids into solvents. A further type includes the catalytic processes such as hydrogenation, in which there is a basic gas-liquid mass transfer to be satisfied, but in addition, effective mixing and shear rate on the catalyst particle fluid film as well as degradation must be considered.

IV. GAS-LIQUID CONTACTING

Many times a specification calls for a fluid mixer to produce a "good dispersion" of so many computational fluid mixing (CFM), of gas into a given volume of liquid. Actually, there are very few applications in which dispersion of gas-liquid is the ultimate process requirement. Usually there is a mass transfer requirement involved, and the role of a mixer to provide a certain mass transfer

coefficient $K_G a$ can entirely supercede any requirement for a particular type of visual description of the gas–liquid dispersion. In general, linear gas superficial velocity, normally given the symbol F , in feet per second, is based on dividing the tank cross-sectional area by the flow of gas at the temperature and pressure of the gas at the midpoint of the tank. This quantity is very basic both in the scaleup correlation and in predicting the power imparted to the liquid by the gas stream.

It is characteristic that this ratio F increases on scaleup, since if we maintain equal volumes of gas per volume of liquid per time on scaleup, which is necessary to provide the same stoichiometric percentage of gas absorbed from the gas phase, then the linear velocity increases directly proportional to the depth of the tank.

While the variables are many and complex, in a general concept, if the power in the tank is equal to the energy provided by the gas stream, we will get a gas-controlled flow pattern. This has different characteristic coefficients of the mass transfer rate than the case where the mixer horsepower is three or more times higher than the gas power. For radial flow impellers, this factor of three will provide a mixer-controlled flow pattern, which again, has different exponents on the correlating equation for mass transfer coefficient $K_G a$ or $K_L a$. To drive the gas down to the bottom of the tank, below the sparge ring, the power level must be on the order of 5–10 times higher than the gas power level.

For axial flow impellers, the ratio of mixer power to gas stream power for a mixer-controlled flow pattern is approximately 8–10. This means that radial flow impellers are more commonly used for gas–liquid dispersion than axial flow impellers.

Figure 20 gives a typical curve for the effect of gas velocity and power level on mass transfer coefficient $K_G a$. In a given application, knowledge of the required gas ab-

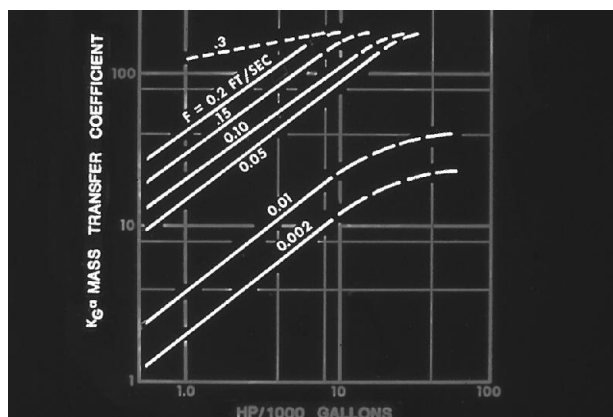


FIGURE 20 Typical correlation of gas–liquid mass transfer coefficient $K_G a$ as a function of impeller power and superficial gas velocity.

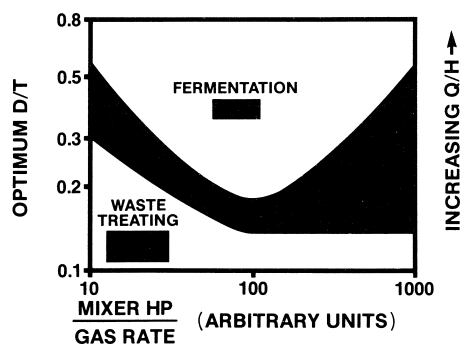


FIGURE 21 Schematic representation of optimum D/T as a function of flow of gas compared to mixer horsepower input. Shaded area is optimum D/T . Two industrial examples, fermentation and aeration of biological waste, are shown.

sorption rate and the partial pressure in the incoming–outgoing gas stream, coupled with an estimate of the equilibrium partial pressure of gas related to the dissolved gas in the liquid, allows the calculation of the average concentration driving force and then the mass transfer coefficient $K_G a$ when needed to provide that mass transfer rate. This then allows the mixer to be chosen for that particular combination. It is typical to try different gas rates, different tank shapes, or perhaps different head pressures to see the effects on the mixer design and the cost for process optimization.

Another consideration is the optimum flow to fluid shear ratio involved for gas–liquid dispersion. Figure 21 shows the optimum D/T for different combinations for gas flow and mixer power level in conceptual form. At the left edge of the curve, where gas rates are high and power levels are low, large D/T values are desired to produce high flow and low shear rates. In the middle of the graph, which is more common, where the gas flow pattern is controlled by the mixer, desired D/T values are very small (on the order of 0.15–0.2). At the far right-hand side of the graph, we have a mixer power level greater than 10 times the gas power level, and it makes very little difference what ratio of flow-to-fluid shear rate we have, as shown by the effect of D/T . This relationship shows the difficulty in comparing impellers in gas–liquid mass transfer systems, because the comparison of fluid shear and fluid flow requires a knowledge of the mixer power to gas flow ratio. In addition, in a process such as fermentation, where there are certain maximum shear rates possible without damaging the organism, the D/T chosen for the process may not be the optimum for the gas–liquid mass transfer step, and correlations must be available for the effect of D/T ratio on mass transfer coefficient to complete design of those kinds of processes.

Scaleup is normally based on the fact that the correlation of $K_G a$ versus power per unit volume and superficial gas velocity is the same for both pilot and full-scale tanks.

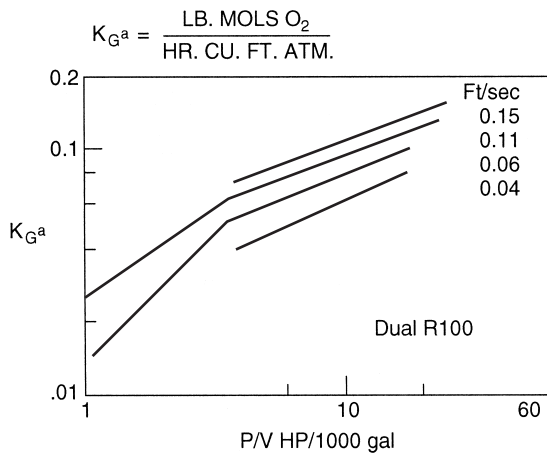


FIGURE 22

This allows the calculation of full-scale mixers when pilot plant data is available in that particular fluid system.

The curve shown in Fig. 22, for an R100 impeller illustrates that there is a break point in the relationship with K_{Ga} versus the power level at the point where the power of the mixer is approximately three times the power in the expanding gas stream. The power per unit volume for an expanding gas stream at pressures from 1 to 100 psi can be expressed by the equation P/V (HP/1000 gal) = $15F$ (ft/sec). The A315 impeller, Fig. 23, is able to visually disperse gas to a ratio of about 1 to 1 in expanding gas power and mixer power level. It does not have a break point in the curve, although slopes are somewhat different than those in Fig. 22.

A comparison of the curves is such that in some areas the A315 has a somewhat better mass transfer and in other areas the R100 has a better mass transfer performance.

The large difference in the A315, however, is more in its blending ability compared to the R100, so that the blend

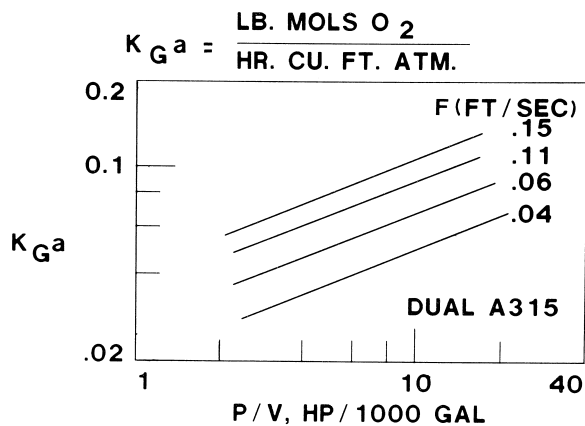


FIGURE 23 Effect of power per unit volume, P/V , and superficial gas velocity, F , on the mass transfer coefficient K_{Ga} , for dual A315 impellers.

time in a large mixing vessel equipped with A315 impellers will be about one-third that of the blend time in the same vessel equipped with R100 impellers. Blending is relatively long on full scale compared to pilot scale, so the improvement in blending characteristics on full scale can lead to a much more uniform blending condition. Many fermentations are responsive to improved blending and this is another factor in addition to the requirement of gas-liquid mass transfer that exists in many fermentation systems as well as in other gas-liquid operations.

A. Combination of Gas-Liquid and Solids Systems

As mentioned previously, axial flow impellers are typically used for solids suspension. It is also typical to use radial flow impellers for gas-liquid mass transfer. In combination gas-liquid-solid systems, it is more common to use radial flow impellers because the desired power level for mass transfer normally accomplishes solids suspension as well. The less effective flow pattern of the axial flow impeller is not often used in high-uptake-rate systems for industrial mass transfer problems. There is one exception, and that is in the aeration of waste. The uptake rate in biological oxidation systems is on the order of 30 ppm/hr, which is about $\frac{1}{2}$ to $\frac{1}{10}$ the rate that may be required in industrial processes. In waste treatment, surface aerators typically use axial flow impellers, and there are many types of draft tube aerators that use axial flow impellers in a draft tube. The gas rates are such that the axial flow characteristic of the impeller can drive the gas to whatever depth is required and provide a very effective type of mass transfer unit.

B. Effect of Gas Rate on Power Consumption

At a given mixer speed, there is a reduction in the horsepower of a mixer when gas is added to the system, and normally the horsepower decreases somewhat proportional to the increase in gas velocity. Figure 24 shows a typical curve, but there are many other variables that affect the location of the curve markedly. This brings up a key point for industrial design. If the mixer is to be run both with the gas off and on in the process, then an interlock is used to prevent gas-off operation or change the mixer speed to prevent overloading, or else the mixer must be capable of transmitting power and torque possibly two, three, or four times higher than is needed during the actual process step. This often is solved by a two-speed motor, which allows a lower speed to be used when the gas is off, compared to the normal speed at processing conditions.

A new impeller is now being used for gas-liquid contacting, call the Smith Turbine. It is a radial flow turbine with blades as shown in Fig. 25. It is rotated in the concave

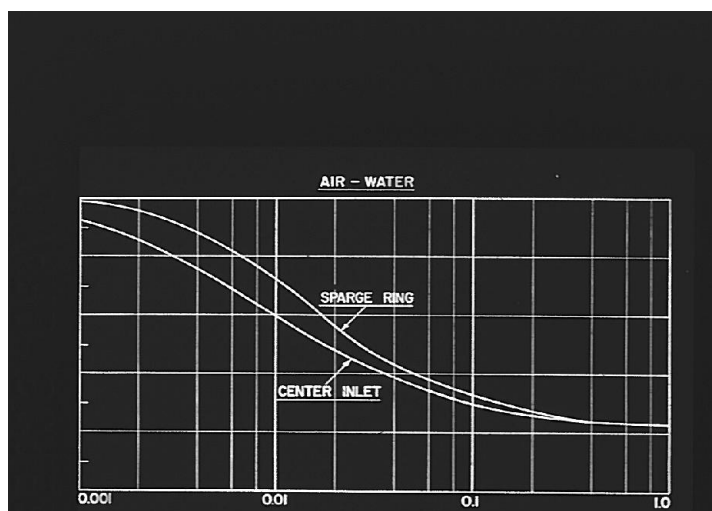


FIGURE 24 Typical plot of K factor, ratio of horsepower with gas-on to horsepower with gas-off at constant speed, as a function of superficial gas velocity with two different gas inlets.

direction. It has the characteristic of giving the same mass transfer as the radial flat blade turbine shown in Fig. 1a but does not drop off in power as much as the radial flow turbine does. Figure 25 shows the K factor which relates the power drawn when the gas is on to the power drawn in the ungassed liquid. This means that variable speed drives do not have to be used many times to keep the desired power in the gased condition.

V. LIQUID-LIQUID CONTACTING

A. Emulsions

There is a large class of processes where the final product is an emulsion. It includes homogenized milk, shampoos, polishing compounds, and some types of medical preparations. A key factor is the chemistry of the product, to ensure that the emulsion remains stable over a desired product shelf-life, when produced properly by the fluid mixer. There are numerous correlations in the literature on drop size in a two-phase liquid-liquid system, relative to fluid properties and mixer variables. However, most

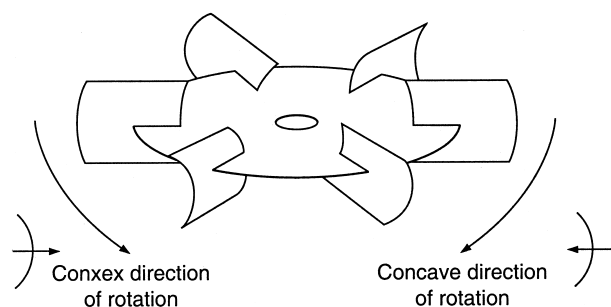


FIGURE 25 Radial flow turbine with blades.

of these have been done with pure liquid components, and do not apply to the complicated chemicals used industrially. Small amounts of surface active agents make dramatic differences in emulsion characteristics, so it is not usually possible to calculate in advance the mixer needed to provide the particular type of emulsion particle size. Therefore, test work is normally required where conditions required in the pilot plant are evaluated for scaleup.

Large tanks have a longer blend time and a much greater variety of shear rates than small tanks, therefore, emulsion characteristics on full scale are difficult to predict. Usually, the pilot plant work is aimed at trying to elaborate the key role of maximum impeller zone shear rate, average impeller zone shear rate, and general circulation rate and velocity out in the main part of the tank. If these can be even qualitatively determined, scaleup to full scale can be done with reliability. There are a variety of mixers used in these processes. For various types of emulsion polymerization, it is typical to use axial flow impellers because the shear rate requirements do not demand the use of radial flow impellers. Getting into the other end of the spectrum, where extremely high shear mixers are needed, various kinds of radial flow blades, usually with very narrow width blades, allow speeds to go up to 1000 or 2000 rpm giving very intense shear rates that are needed for many types of emulsion processes.

B. Liquid Extraction

Figure 26 shows a typical curve relating the performance of many kinds of mixing devices for liquid extraction. The main advantage of using mixing or some type of mechanical energy, compared to packed plate or spray

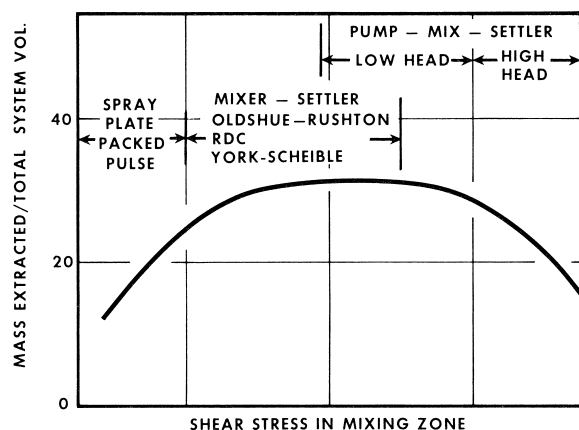


FIGURE 26 Illustration of optimum shear stress in a mixing zone of various types of countercurrent liquid-liquid extraction columns.

columns, is the ability to get a smaller volume for the same degree of extraction. However, if an attempt is made to use too much energy, then problems of settling characteristics are encountered, and this negates the advantages of the mixed system many times. In the mining industry, it is quite typical to use mixer settlers. These usually involve an extraction step, a scrubbing step, and then a stripping step. Usually the requirement is for only one or two stages in each of these areas with the use of very selective ion exchange chemicals in the system. To eliminate interstage pumps a pump-mixer is used in which some of the head component of the impeller is converted to a static head so that fluids can be pumped against small static heads in the mixers and settlers of the whole train. This has worked well in many applications, although there is a potential problem that the conditions required for effective pumping are not optimum for the mixing that is required in the mixing stage, and there may be some design parameters that are difficult to satisfy in the systems.

The other area is the countercurrent liquid-liquid extraction system, shown in Fig. 27, using mixer stages separated by stationary horizontal discs. These have the advantage of only one interface for settling to occur, plus the fact that solids can be handled in one or both phases. Also, all the principals of fluid mixing can be used to design an effective transfer system. The design procedure is also based on the $K_L a$ concept, discussed in Section IV, and allows the calculation of reliable full-scale performance, based on pilot plant work, often done in a laboratory column about 6 in. in diameter.

One of the key variables to be studied in the pilot plant is the effect of turndown ratio, which is the ratio of flow to the design flow through the column, so that predictions can be made of performance during reduced throughput during certain parts of the plant processing startup.

VI. BLENDING

A. Low-Viscosity Blending

Low-viscosity blending involves evaluation of the degree of uniformity required and the operating cycle. There is a difference in performance, depending on whether the materials to be blended are added continuously and

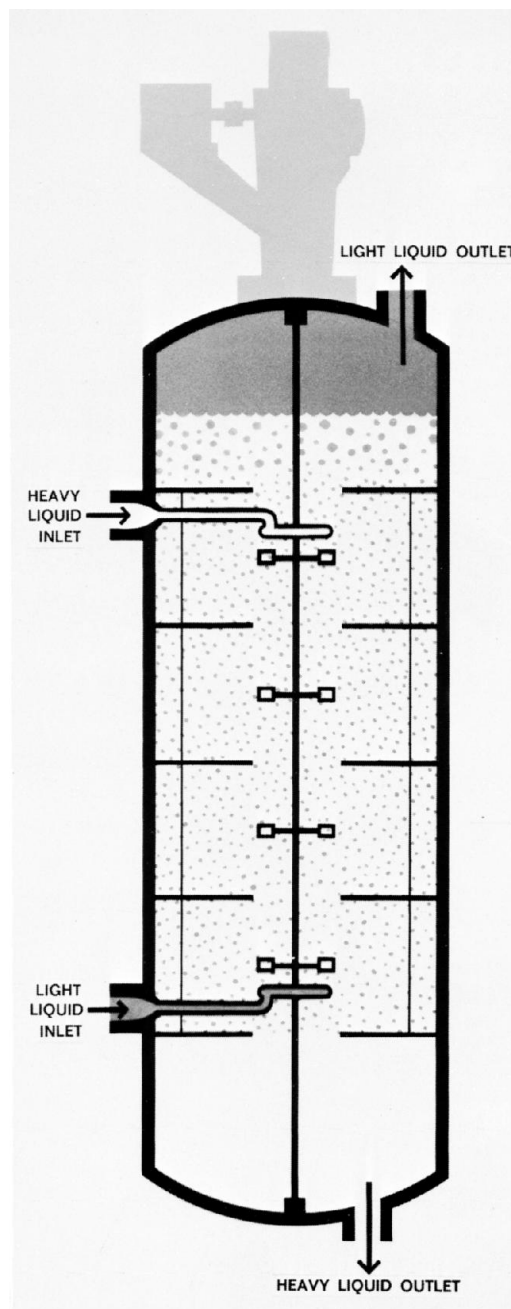


FIGURE 27 Typical countercurrent liquid-liquid extraction column with mixing phases: Oldshue/Rushton column illustrated.

uniformly into the tank, with the tank originally in motion or whether the tank has become stratified during the filling application, and mixing must be accomplished with a stratified liquid level situation. In general, blend time is reduced at constant mixer power with larger D/T ratios. The exponent on D/T with blend time is approximately -1.5 , with the range observed experimentally of from 0.5 to 3.0. This leads to the fact that larger impellers running at slow speeds require less power than a small mixer running at high speed for the same blend time. In that case, there is an evaluation needed which relates the capital cost of the equipment, represented by the torque required in the mixer drive which is usually greatest for the big impeller, versus the cost of horsepower, which is usually greatest for the small impeller. This leads to the concept of optimization of the economics of a particular process. In all cases, at least two or three mixers must be selected for the same blend time with different power and impeller diameter to carry out this evaluation.

Table V gives a typical values for estimation purposes of blending horsepower required for various low- and medium-viscosity situations. Mixers may be either top entering or side entering. Again, a side-entering mixer requires more power and less capital dollars, and this must be evaluated in looking at practical equipment. Side-entering mixers have a stuffing box or mechanical seal and are limited for use on materials that are naturally lubricating, noncorrosive, or nonabrasive.

B. High-Viscosity Blending

Blending of high-viscosity materials, which are almost always pseudo-plastic, involves a different concept. The degree of pseudo-plasticity is determined by the exponent n in the equation

$$\text{shear stress} = K(\text{shear rate})^n$$

with value 1 for Newtonian fluids and a value less than 1 representing the degree of decrease of viscosity with an increase in shear rate. For very viscous materials (on the order of 50 m cP and higher), the helical impeller

TABLE V Motor Horsepower for Estimating Purposes for Blending Purposes^a

Blend time θ (min)	H.P. ^b			
	100	250	500	1000
6	5	7.5	10	15
12	3	5	7.5	10
30	1.5	2	3	5

^a 15,000 gal tank; axial flow impeller, $D/T = \frac{1}{3}$, $Z/T = 1$; $C/D = 1$.

^b For viscosities in centipoises.

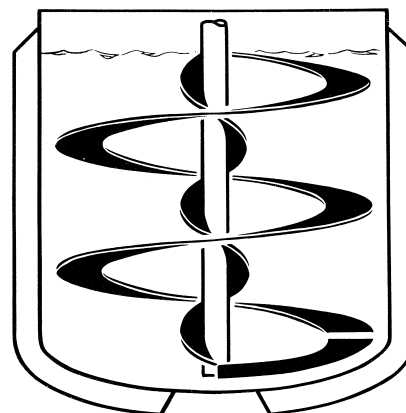


FIGURE 28 Typical helical flow impeller for high-viscosity blending with close clearance to the tank wall.

(Fig. 28) is often used. Many times this is a double helix, in which pumping on the outside is done by the outer flight, while pumping on the inside is done by the inner flight. Reverse rotation, of course, reverses the direction of the flow in the tank. These impellers typically run at about 5–15 rpm and have the unique characteristics that the circulating time and blend time are not a function of the viscosity of the fluid. At a given velocity, there is a certain turnover time for a given Z/T ratio, and changing viscosity does not affect that parameter, nor does the degree of pseudo-plasticity affect it. However, the power is directly proportional to the viscosity at the shear rate of the impeller, and so doubling or tripling the viscosity at the impeller shear rate will cause an increase of power of two or three times, even though circulation time will remain the same. Helical impellers are very effective for macroscale blending, but do not typically have the microscale shear rate required for some types of uniformity requirements or process restraints.

Open impellers, such as the axial flow turbine (Fig. 1a) or the radial flow turbine (Fig. 1b), may also be used in high-viscosity pseudo-plastic fluids. These require a level of power four to five times higher than the helical impeller, but only cost about one-third as much. Another economic comparison is possible to see which is the most effective for a given operation. This higher power level, however, does provide a different level of microscale blending. Occasionally the flow from a blend system with a helical impeller will be passed through a mechanical type of line blender, which imparts a higher level of microscale mixing.

C. Side-Entering Mixers

Figure 29 shows the importance of orientation on side-entering mixers on low-viscosity systems. The mixer must be inclined about 7° from the tank diameter, to ensure a

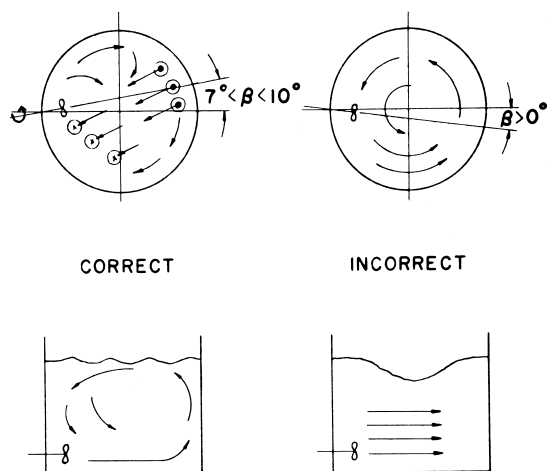


FIGURE 29 Typical orientation of side-entering mixer in large petroleum storage tanks.

top-to-bottom flow pattern. However, even when this is done, there still are some relatively stagnant areas of the tank, and side-entering mixers are not usually satisfactory when solid suspension is a critical factor. As discussed previously, larger-diameter impellers at slower speeds require less horsepower, and so there can be an economic evaluation of the power versus capital equipment cost for various types of side-entering mixers and a given blending process. Typical applications are crude oil tanks, gasoline tanks, and paper stock. In addition, they are used for various kinds of process applications where the advantages are considerable over the use of a conventional top-entering mixer.

VII. FLUID MOTION

Many times the objective is to provide a pumping action throughout the tank. The pumping capacity of impellers can be measured by photographic techniques, hot wire or hot film velocity meters, or laser Doppler velocity meters. There is no generally agreed upon definition of the discharge areas for impellers, so that the primary pumping capacity of mixing impellers varies somewhat, depending on the definition used for discharge area. There is considerable entrainment of fluid in the tank, due to the jet action of the flow from the impeller. Figure 30 shows the increase in total flow in the tank at various D/T ratios. This also indicates that at about 0.6 D/T ratio further increases in total flow in the tank are difficult to achieve, since there is no more entraining action of the impeller in the total system.

The pumping capacity of a mixing impeller is specified by either the flow from the impeller or the total flow of the tank. Flow varies for any impeller as the speed and diameter cubed. Table VI gives some for constants in the equation $Q = KND^3$ for various impeller types. The radial

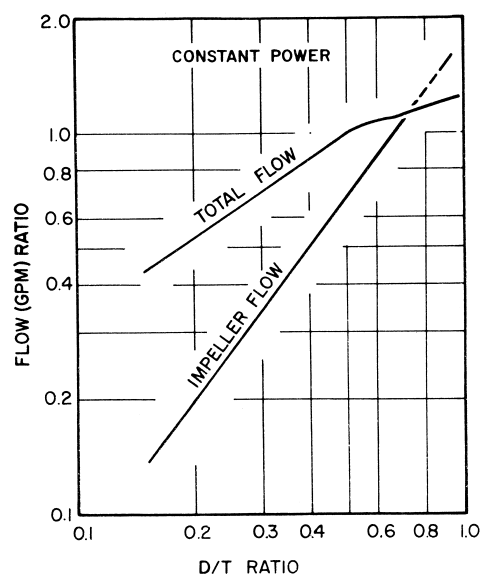


FIGURE 30 Schematic illustration of total flow in mixing tank as compared to impeller flow.

flow impeller has essentially less flow and higher shear rates than does the axial flow impeller type.

If the impeller is required to pump against a static head or a friction head within the channel of the mixing tank, then there must be a series of head flow curves developed, (Fig. 31) for the impeller being used. This is a function of the clearance between a radial impeller and a horizontal baffle. The hole in it allows the flow to come into the impeller zone but not circulate back, or the clearance of an axial impeller in a draft tube (Fig. 32). The operating point, then, will be the intersection of the impeller head flow curve and the system head flow curve. Draft tube circulators have the advantage of giving the highest flow in the annulus for a given level of power or requiring the least power to provide a given flow of the annulus. When pumping down the draft tube, the flow in the annulus must equal the settling velocity of the particles, and the total flow can be calculated on that basis. In practice, the flow coming up the annulus is not a uniform flat velocity profile; so that additional total flow is needed because of the nonuniform distribution of the upward axial velocity to the annulus. Pumping down the draft tube allows the tank bottom to be flat or have very small conical fillets at the sidewalls.

Pumping up the draft tube requires that the solids are to be suspended in the draft tube with a much lower total

TABLE VI Constant in Flow versus Speed and Diameter of Various Mixing Impellers

Figure	K
1a	0.8
1b	0.6

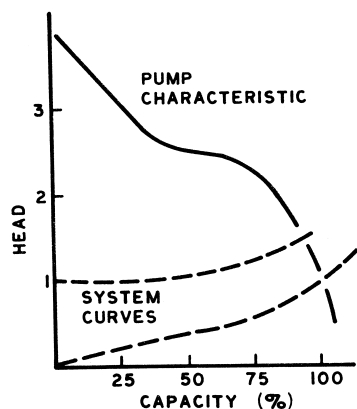


FIGURE 31 Typical head flow curve for mixing impeller and draft tube with corresponding system curves.

flow, and also power, and then make their own way down the outside of the annulus coming into the bottom of the draft tube again. This means that the bottom of the tank must usually have a steep cone, and suitable flares and baffles must be added to the draft tube bottom so that the flow comes up in a uniform fashion for proper efficiency.

When using a draft tube, the back flow possibility in the center of the impeller requires the use of a large-diameter hub. This is not normally desirable in fluidfoil impellers used in open tanks. The system head for a draft tube circulator is a function primarily of the design of the entrance and exit of the draft tube, and considerable work has been

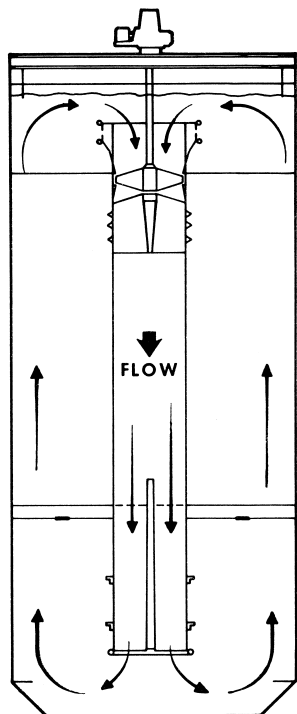


FIGURE 32 Typical axial flow impeller and draft tube.

done on the proper design and flaring of these tubes for special applications. The main use of draft tube circulators has been in precipitators and crystallizers. A further requirement is that the liquid level be relatively uniform in depth above the top of the draft tube, which means that variable liquid levels are not practical with draft tube systems. In addition, slots are often provided at the bottom of the draft tube, so that should a power failure occur and solids settle at the bottom of the tank, flow can be passed through these slots and scrub out particles at the bottom of the tank for resuspension.

Sometimes it is desired to have a large working area in a tank where, for example, a conveyor belt containing car bodies can be passed through for electrostatic painting. One way to accomplish this is to put a series of propeller mixers in a side arm of the long side of the tank, so that the flow is directed into the middle zone, but there are no mixer shafts or impellers in the center to impede the flow of the parts through the equipment.

VIII. HEAT TRANSFER

Another area for pumping consideration is heat transfer. The only sources of turbulence provided in heat transfer are flow around the boundary layer of a jacketed tank and around a helical coil or vertical tubes. There are several good heat transfer correlations available, and most of them have fairly common exponents on the correlation of the Nusselt number hD/k . This is correlated with the Reynolds number $ND^2\rho/\mu$ and the Prandtl number $Cp\mu/k$ plus other geometric ratios. The exponential slope on the effect of power on heat transfer coefficient is very low (on the order of 0.2). This means that most heat transfer design involves determining the mixer required for just establishing forced convection through the tank, and usually not going beyond that point if heat transfer is the main requirement. If other requirements are present which indicate a high horsepower level, then advantage can be taken of these higher power levels by use of the 0.2 exponent. However, if it is desired to increase the heat transfer capacity of a mixing tank, it is normally done by increasing or changing the heat transfer surface, since very little can be done by changing the mixer power level. Figure 33 gives a good working correlation for the effect of viscosity on both heating and cooling coefficients for helical coil systems. Jacketed tanks have values about two-thirds of those in Fig. 33. This is the mixer side coefficient only, and it holds for organic materials. The heat transfer coefficient for aqueous materials is higher than the value shown in Fig. 33. Bear in mind that the overall coefficient is made up of other factors, including the coefficient on the inside of the tube or jacket, as well as the thermal conductivity value of the heat transfer surface.

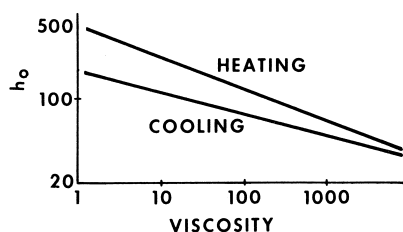


FIGURE 33 Practical heat transfer coefficients for use in estimating with helical coils and vertical tubes.

IX. CONTINUOUS FLOW

A mixing tank has a variety of residence times. The definition of perfect mixing requires that one particle leave in 0 time and one particle stay in forever. Curves shown in Fig. 34; developed by McMullen and Weber, show the percentage of material that is in the tank for various lengths of time. To provide good mixing in a system but avoid the detrimental effect of a variety of residence times, multiple staging can be used. This curve shows, for example, that if the total residence time in a tank were 60 min, then at the end of 30 min, 33% of the material is already gone and 67% of the material is still there. Out at the very long residence time, there is still a small amount of material that stays in an infinitely long length of time. This means that processes involving pharmaceuticals or food products must take into account that small contaminants or mutants may stay in the system for a very long time and can cause problems in yield and productivity.

Another purpose of a mixing tank is to dampen out fluctuations. A mixing tank cannot change the frequency of fluctuations but can dampen the amplitude. As a general principle, a residence time equal to the cycle time of the fluctuations will cause the amplitude to be dampened by about a factor of six.

For any chemical reaction of an order greater than zero, the process takes longer in a continuous flow tank than it

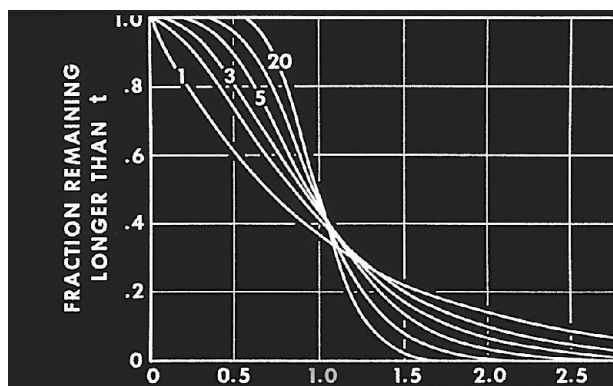


FIGURE 34 Curves based on perfect mixing in each compartment of the multistage compartment system, showing percentage material retained for various lengths of time in continuous flow.

does in a corresponding batch tank. An infinity of mixing stages is equivalent to a batch tank or to a plug flow reactor. Usually, however, 5, 10, or 20 stages are sufficient to give a good efficient reaction time and to possess the advantages of continuous flow compared to the reaction time in a batch system.

A. Inline Mixers

Mixers in a flowing pipeline are of two general types, one utilizing static elements and the other using a rotating impeller.

A static inline mixer is essentially a device that provides transverse uniformity and not longitudinal or time-interval blending. Hence, if a particle in Fig. 35 is ever to catch up with another particle behind it, there must be a tank volume such that the first particle can remain until the latter one catches up with it.

There are two kinds of static mixers. One type has helical elements that twist the fluid, and another set of elements that cut the fluid, divide it, and twist it again. The twisting and cutting is continued until the production and scaleup uniformity is achieved. This is useful in viscous fluids.

Attempts to use these kinds of devices on low-viscosity materials showed that the flows did not twist and curl in quite that same fashion. In the low-viscosity region, pressure drop is a key factor. The second type of static mixer gets pressure drop through controlled channels, different types of static elements, as well as random placement of baffles, blades, orifices, or other devices inside the pipeline.

Mechanical inline mixers have a relatively high-speed impeller, rotating in a small volume, usually on the order of $\frac{1}{4}$ gal to perhaps 50 or 60 gal. Obviously, with a big enough tank, you then have a system that really does not fit in the pipe-line itself. Usually, the flow is directed through two stages, the flow comes in the bottom of the container, flows up through a hole in a static plate into a stage divider, and then flows in the second impeller. The power is such that

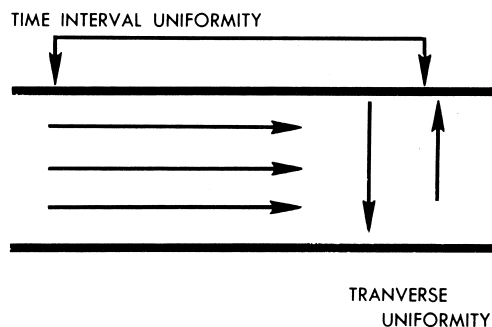


FIGURE 35 Pipeline flow showing that time-interval mixing normally must have a volume for retention time, compared to radial flow with usual static mixer elements.

the flow pattern is completely disrupted, so the pressure drop to these units is at least one velocity head. The rpm can be adjusted to achieve almost any required level of dispersion for contacting.

X. PILOT PLANT PROCEDURES

Pilot planting involves gathering sufficient information from model runs so that the major controlling factors in the process are understood for a suitable scaleup analysis.

The heart of the pilot plant study normally involves varying the speed over two or three steps with a given impeller diameter. The analysis is done on a chart, shown in Fig. 36. The process result is plotted on a log-log curve as a function of the power applied by the impeller. This, of course, implies that a quantitative process result is available, such as a process yield, a mass transfer absorption rate, or some other type of quantitative measure. The slope of the line reveals much information about likely controlling factors. A relatively high slope (0.5–0.8) is most likely caused by a controlling gas–liquid mass transfer step. A slope of 0, is usually caused by a chemical reaction, and a further increase of power is not reflected in the process improvement. Point A indicates where blend time has been satisfied, and further reductions of blend time do not improve the process performance. Intermediate slopes on the order of 0.1–0.4, do not indicate exactly which mechanism is the major one. Possibilities are shear rate factors, blend time requirements, or other types of possibilities.

To further sort out the effect of mixing, it is usually desirable to vary the impeller diameter. For example, if a 100-mm impeller had been used in a 300-mm diameter tank for the original runs, and if it were thought that pumping capacity would be more helpful in fluid shear rate, a series of runs with 125- or 150-mm diameter im-

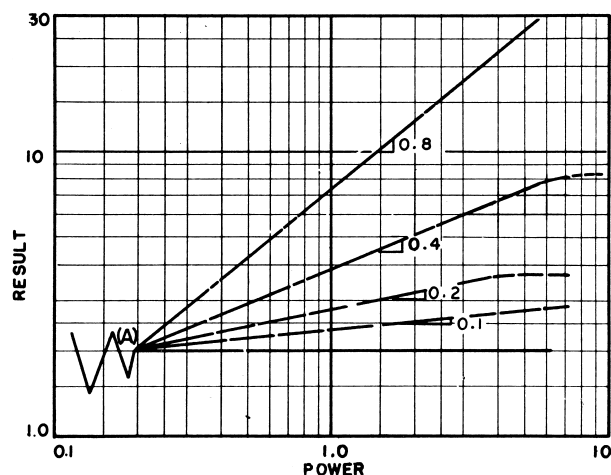


FIGURE 36 Typical plot of a given process result as a function of mixer power level in a pilot plant study.

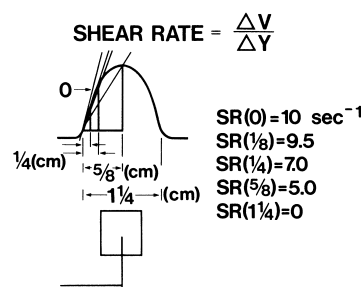


FIGURE 37 Schematic illustration that the macroscale shear rate around the impeller is a function of the size of the fluid element of interest.

peller would be appropriate. On the other hand, if it were thought that fluid shear was more important, then runs with a 50- or 75-mm impeller would be indicated.

If separation of the microscale mixing phenomenon from the macroscale mixing phenomenon is desired, then it is necessary to systematically vary the ratio of blade width to blade diameter.

There is a minimum size pilot tank. Referring now to Fig. 37, the shear rate at the boundary layer of the impeller jet in the tank has approximately a value of 10 in this example. The impeller is approximately 1 cm in blade width. The shear rate across a $\frac{1}{8}$ cm is about 9.5, shear rate across a $\frac{1}{4}$ cm is 7.5, and the shear rate across a $\frac{1}{2}$ centimeter is 5, and is the average shear rate. The shear rate across the entire blade 1 cm wide is 0, since it has the same velocity on both sides of the impeller blade. Thus, a particle of 1 cm size would have a zero shear rate, while a particle having a $1 \mu\text{m}$ size would have a shear rate of 10. This leads to the general rule that the impeller blade must be at least three times larger in physical dimension than the biggest particle that is desired to disperse, react, or coalesce. In practice, this indicates that most gas–liquid processes should be done in tanks at least 12 in. in diameter, while most viscous and pseudo-plastic materials should probably be handled in tanks from 12 to 18 in. in diameter. Homogenous chemical reactions could be carried out in a thimble, if desired, since there is no problem getting the scale of the molecule to be smaller than the scale of an impeller blade, even a small laboratory size.

It is usually desirable to either measure or calculate horsepower, and there are several methods by which this can be done. One is to have impellers calibrated by the manufacturer, which provides a curve of power versus speed. By using suitable factors for judging viscosity and gas flow, power in the batch can be estimated as a function of the impeller speed. Another possibility is to place the impeller on a trunion bearing mounting, in which the motor is held stationary by a pulley arm, and the force required is measured on a scale. Another method involves the use of strain gauges, which measure either the elongation on the surface of a shaft or the changes in conductivity

or reluctance with various kinds of electrical signals. It is possible today to use micro-sized amplifiers that rotate with the shaft and feed a signal through the slip rings with very little loss in accuracy.

In general, a large mixing tank has a much longer circulation time and a much higher maximum macroscale impeller shear rate than does a small tank. In addition, it has a greater variety of shear rates than does a small tank. This means that a small tank can be changed in its performance compared to a big tank by using a nongeometric approach to the design of the mixer. There are usually two extremes of pilot plant objectives. One involves the use of a more-or-less standard impeller geometry in small scale, and attempts to determine the maximum efficiency of the process on that scale. Estimates on a full-scale performance must be modified because the big tank is different in many regards, which may have beneficial or detrimental effects on the process.

The other approach looks at either existing equipment in the plant or a probable design of a full-scale device. How can this be modeled in a pilot plant? This usually involves using narrow-blade impellers and/or small-diameter impellers to more closely decrease the blend time and increase the shear rate over what might usually occur when geometric similarity is used in a pilot plant.

In addition, the variety of shear rates in a big tank means that for bubble or droplet dispersion requirements, the big tank will have a different distribution of bubble sizes than the small tank. This can be very important in such areas as polymerization and particle size analysis.

A. Step #1—What to Do First

First ask yourself if there is any role for fluid shear stresses in determining and obtaining the desired process result. About half of the time the answer will likely be no. That is the percentage of mixing processes where fluid shear stresses either have no effect or seem to have no effect on the process result. In these cases, mixer design can be based on pumping capacity, blend time, velocities and other matters of that nature. Impeller type location and other geometric variables are major factors in these types of processes.

However, if the answer to this first question is yes; there is an effect of fluid shear stresses on the process, then there needs to have a second question asked. Is it at the micro- or macroscale that the process participants are involved? And, of course, it may be both.

B. Scaleup/Scaledown

Table III shows what happens to many of the variables on scale up. A summary of this is that blend time typically increases and the standard deviation of circulation times

around the mean circulation times also normally increases. The quantitative effect depends somewhat on the degree of uniformity required and the blend time being considered.

As a general rule, the operating speed of the mixer tends to go down, while the peripheral speed of the impeller tends to go up. The speed of the mixer is related to the average impeller zone macroscale shear and thus typically goes down in scaleup while the impeller peripheral speed is often related to the maximum impeller zone macroscale shear rate, see Fig. 5. Out in the rest of the tank (away from the impeller) there another spectrum of shear rates which typically is about a factor of 10 lower than the average impeller zone shear rate. These particular impeller zone shear rates tend to decrease on scaleup.

The microscale environment tends to have a power per unit volume of dissipation around the impeller about 100 times higher than it is in the rest of the tank more or less regardless of the tank size. Thus, the magnitudes of these quantities can be quite similar. This brings up another consideration in the following paragraph.

C. Shear Rate Magnitude and Total Shear Work

Shear stresses and their origin from shear rates (shown in Table VII) gives the magnitude of the shear stress environment that the process participants see. The time they are exposed to that magnitude is a major factor in the process result. For example, it may take a minimum shear stress magnitude to create a certain size particle. However, the ultimate distribution of particle sizes may well relate to the length of time that a particle is exposed to that shear rate. The product of shear stress and time determines what is likely to happen to the process. This obviously is a matter of the spectrum of shear stresses throughout the tank and the statistical distribution of circulation times that particles have going through these zones.

With constant viscosity between the model and the prototype and/or a constant change in viscosity to the process during a batch operation, we can substitute shear rate for shear stress and the product of shear rate times the time is a dimensionless number. Considerable progress is being made toward calculating the velocities, shear rates, and circulating times in mixing vessels, and suitable models and calculations could be made to model these effects in more quantitative detail both on a point-by-point basis and at an overall vessel average. What still is challenging, however,

TABLE VII

Fluid Shear Stress	= Viscosity	Fluid Shear Rate

is that it is not usually known what effect these particular properties will have on the process participants in a given process and, thus, it is usually necessary to measure the process result either full scale in the plant, or in smaller size systems in pilot plant or laboratory.

To summarize the situation, geometric similarity controls no mixing variable whatsoever. The question is does that make a difference to the process. In the portion following, we will take a look at the ten basic mixing technology classifications and see what effect these considerations might have. Added to this is the fact that most industrial mixing processes involve two or more of the ten mixing technological classifications and so their interaction between those technology classification parameters must be considered to give the overall performance of the mixing process.

D. What to Do in the Pilot Plant

There are several considerations to bear in mind when planning a pilot plant program.

1. The pilot tank is blending rich while full scale tanks are blending poor. This means that relatively inefficient blending impellers are needed in the pilot

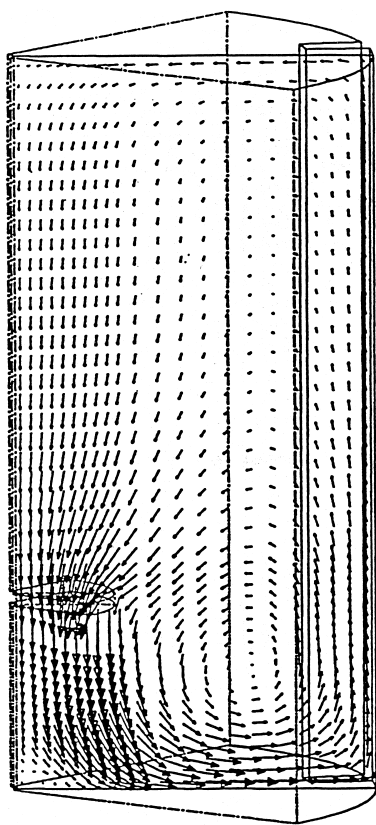


FIGURE 38 Velocity vectors for an A310 impeller.

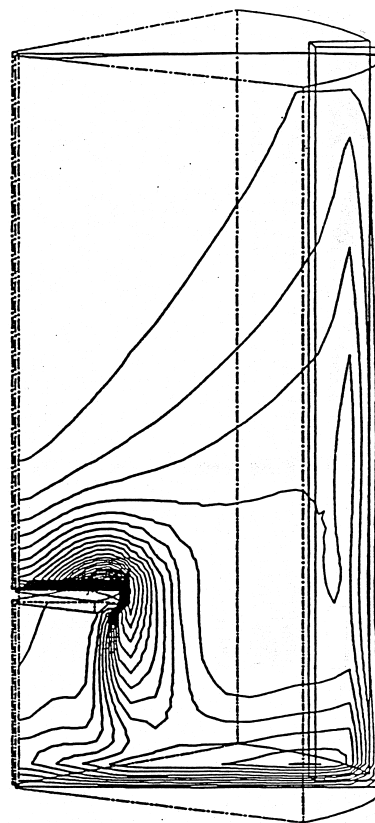


FIGURE 39 Contours of kinetic energy of turbulence.

- plant to correspond to the blending efficient impellers used in the plant.
2. One technique to make the pilot plant unit more similar to the plant scale unit is to use impellers of relatively narrow blade width compared to their traditional blade widths used with commercial impellers in the plant. This is purposely reducing the blending performance and improving the shear rate performance in the pilot plant by using impellers of relatively narrow blade width. The blade width cannot be so small that it gets out of proportion to the process participant particles.
3. Always bear in mind the qualitative relationship with viscosity is that the full-scale tank will appear to be less viscous than the pilot plant tank, somewhere in the range of a factor of 10–50.
4. If it appears that upon a qualitative examination that the role of circulation time, blend time, and shear rate may not be important to the process on scaleup, then go ahead and use geometric similarity on the pilot plant study and all of these differences noted above will play no part in the results of the scaleup prediction. It may be that there are compensating effects that while circulation time becomes longer and shear rates become larger, there is a compensating effect that makes the process result satisfactory.

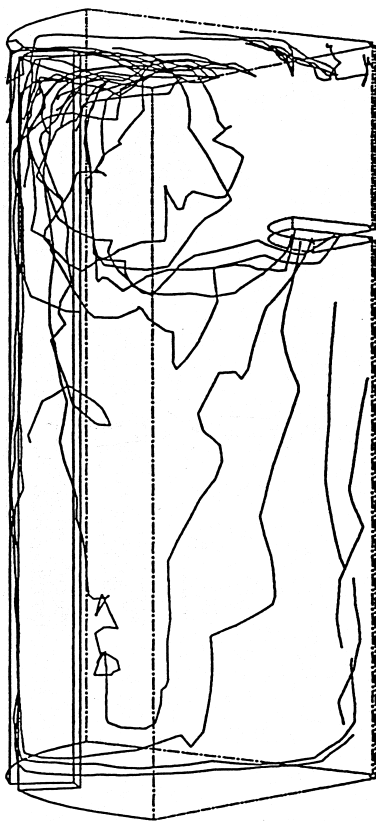


FIGURE 40 A particle trajectory approach with neutral buoyancy particles.

XI. COMPUTATIONAL FLUID DYNAMICS

There are several software programs that are available to model flow patterns of mixing tanks. They allow the prediction of flow patterns based on certain boundary conditions. The most reliable models use accurate fluid mechanics data generated for the impellers in question and a reasonable number of modeling cells to give the overall tank flow pattern. These flow patterns can give velocities, streamlines, and localized kinetic energy values for the systems. Their main use at the present time is to look at the effect of making changes in mixing variables based on doing certain things to the mixing process. These programs can model velocity, shear rates, and kinetic energy, but probably cannot adapt to the actual chemistry of diffusion or mass transfer kinetics of actual industrial process at the present time.

Relatively uncomplicated transparent tank studies with tracer fluids or particles can give a similar feel for the

overall flow pattern. It is important that a careful balance be made between the time and expense of calculating these flow patterns with computational fluid dynamics compared to their applicability to an actual industrial process. The future of computational fluid dynamics appears very encouraging and a reasonable amount of time and effort placed in this regard can yield immediate results as well as potential for future process evaluation.

Figures 38–40 show some approaches. Figure 38 shows velocity vectors for an A310 impeller. Figure 39 shows contours of kinetic energy of turbulence. Figure 40 uses a particle trajectory approach with neutral buoyancy particles.

Numerical fluid mechanics can define many of the fluid mechanics parameters for an overall reactor system. Many of the models break the mixing tank up into small microcells. Suitable material and mass transfer balances between these cells throughout the reactor are then made. This can involve long and massive computational requirements. Programs are available that can give reasonably acceptable models of experimental data taken in mixing vessels. Modeling the three-dimensional aspect of a flow pattern in a mixing tank can require a large amount of computing power.

SEE ALSO THE FOLLOWING ARTICLES

FLUID DYNAMICS • FLUID DYNAMICS (CHEMICAL ENGINEERING) • FLUID INCLUSIONS • HEAT TRANSFER • REACTORS IN PROCESS ENGINEERING • SOLVENT EXTRACTION

BIBLIOGRAPHY

- Dickey, D. S. (1984). *Chem. Eng.* **91**, 81.
- Mcmullen, R., and Weber, M. (1935). *Chem. Metall. Eng.* **42**, 254–257.
- Nagata, S. (1975). “Mixing Principles and Applications,” Halsted Press, New York.
- Nienow, A. W., Hunt, G., and Buckland, B. C. (1994). *Biotech, Bio Eng.* **44**, No. 10, 1177.
- Oldshue, J. Y. (1996). *Chem. Eng. Prog.* Vol. **92**.
- Oldshue, J. Y. (1980). *Chem. Eng. Prog.* June, pp. 60–64.
- Oldshue, J. Y. (1981). *Chemtech.* Sept., pp. 554–561.
- Oldshue, J. Y. (1981). *Chem. Eng. Prog.* May, pp. 95–98.
- Oldshue, J. Y. (1983). “Fluid Mixing Technology,” McGraw-Hill, New York.
- Patwardhan, A. W., Joshi, J. B. (1999). *Ind. Eng. Chem. Pres.* **38**, 49–80.
- Tatterson, G. B. (1991). *Fluid Mixing and Gas Dispersion in Agitated Tanks*.
- Uhl, V. W., and Grey, J. B. (1966). “Mixing Theory and Practice,” Vols. I, II, and III, Academic Press, New York.



Heat Exchangers

Kenneth J. Bell

Oklahoma State University

- I. Applications in Chemical Engineering
- II. Criteria for Selection
- III. Types of Heat Exchangers
- IV. Basic Heat Exchanger Equations
- V. The Design Process
- VI. Further Developments in Design and Application

GLOSSARY

Condensation Conversion of a vapor to a liquid by removing the latent heat of condensation from the vapor.

Condenser Heat exchanger with the primary function of condensing a vapor by transferring heat to a coolant stream.

Conduction Heat transfer within a substance by molecular motion (and also by electron flow in electrical conductors). The molecular motion may be actual displacement of molecules (the predominant mechanism in gases) or may be collisions between adjacent vibrating molecules (the predominant mechanism in liquids and nonmetallic solids).

Convection Heat transfer within a flowing fluid by physical translation of one element of the fluid (consisting of a very large number of molecules) characterized by one temperature to another part of the flow field at a different temperature. The heat is carried as the internal energy of the molecules.

Fouling Unwanted deposit on heat transfer surface due to sedimentation, crystallization, biological films, cor-

rosion, and/or thermal degradation of organic process fluids.

Heat exchanger Any device that allows two or more fluids at different temperatures to transfer heat from the hotter stream(s) to the colder stream(s). Usually the streams are separated by solid walls, but the streams are allowed to mix in *direct-contact* heat exchangers.

Heat transfer coefficient Ratio of the rate of heat transfer in a heat exchanger (in watts) to the product of the heat transfer area of the heat exchanger (in square meters) and the mean temperature difference between the hot and cold streams (in kelvins). The higher the heat transfer coefficient, the more effective the heat transfer process.

Latent heat transfer Transfer of heat required to bring about a phase change (e.g., condensation or vaporization) in a fluid. (Compare to sensible heat transfer.) Many heat exchangers involve both latent and sensible heat transfer.

Mean temperature difference Effective difference between the temperature of the hot stream and the temperature of the cold stream in a heat exchanger. This is the driving force for the heat transfer process.

Pressure drop Decrease in static pressure of a stream between the entrance and the exit of a heat exchanger.

Sensible heat transfer Transfer of heat required to cause a change of temperature in a fluid. (Compare to latent heat transfer.)

Thermal duty Total amount of heat transferred from one stream to the other in a heat exchanger.

Vaporization Conversion of a liquid to a vapor by adding the latent heat of vaporization to the liquid. “Boiling” is a commonly used synonym but is not as precise.

Vaporizer Heat exchanger with the primary function of vaporizing a liquid by transferring heat from a hot stream. Also termed “reboiler” in some applications.

HEAT EXCHANGERS play an essential role in chemical processing. In the typical process plant, heat exchangers bring the feed streams to the proper temperature for the reactors, provide vapor and liquid reflux streams for the separation and purification steps, and finally cool the products for storage and shipping. But the same types of heat exchangers are used in a wide variety of auxiliary services in process plants and many other places as well; examples include lubricating-oil coolers for all kinds of machinery, compressor intercoolers and aftercoolers for gas pipeline systems, chillers in refrigeration and air-conditioning installations, and vapor generators and condensers in conventional, nuclear, geothermal, and solar thermal power plants. Heat exchangers come in many different configurations and with surface areas ranging from 0.1 to 100,000 m². The selection of type or configuration of heat exchanger is governed by the nature of the streams flowing in the exchanger (e.g., liquid or gas, high or low pressure, high or low temperature) and the service (e.g., heating or cooling, condensing, vaporizing) to be performed. The size of the heat exchanger is governed by the amount of heat to be transferred and the rate of heat transfer, which can vary by several orders of magnitude.

I. APPLICATIONS IN CHEMICAL ENGINEERING

A simple but common heat exchanger application in a chemical process plant is cooling a hot liquid or gas product from the process (called the “process fluid”) to a temperature low enough that it can be safely stored. The coolant is likely to be air or water, which would be heated in the heat exchanger. If none of the fluids involved reach their boiling or condensing temperatures, no phase change occurs, and the process fluid is “sensibly cooled” and the coolant “sensibly heated.” A heat balance relates the inlet and outlet temperatures, the specific heats, and the mass

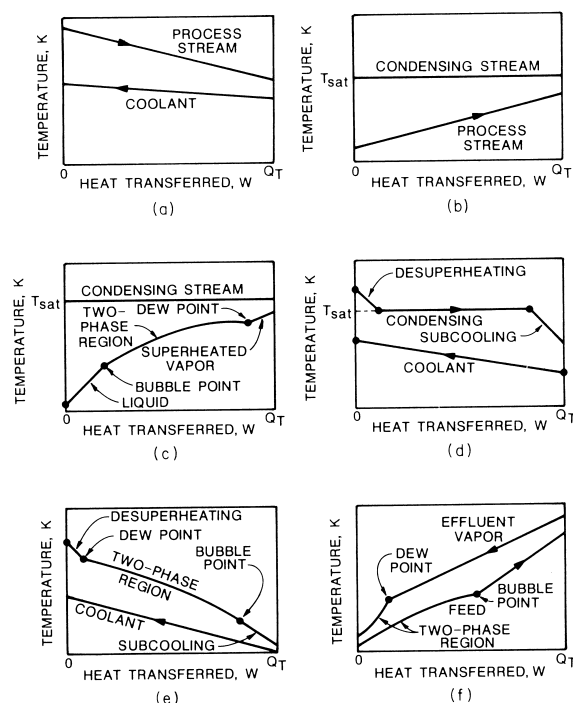


FIGURE 1 Typical temperature profiles for several process heat exchanger applications: (a) product cooler; (b) feed heater with condensing stream; (c) multicomponent feed heater with vaporization and superheating; (d) pure-component product condenser; (e) multicomponent product condenser; (f) typical feed-effluent heat exchanger.

flow rates of the two streams. Since specific heats usually vary little with temperature, the local stream temperatures are linear functions of the heat exchanged between the streams, as shown in Fig. 1a.

If the process fluid needs to be heated (e.g., for feed to a chemical reactor), the hot fluid supplying the heat is likely to be saturated steam at a high enough pressure that the condensing temperature is greater than the final temperature of the process fluid. The heat exchanger is usually designed so that the pressure drop in the condensing steam is negligible compared to the static pressure, so the steam condenses at essentially constant temperature, as shown in Fig. 1b. The heat given up by the steam is the latent heat of condensation and is equal to the sensible heat gained by the process fluid.

A somewhat more complicated situation occurs if a liquid process fluid made up of several components (e.g., crude oil) is to be partially vaporized, possibly as a feed to a distillation column. The liquid heats up sensibly until it reaches the temperature at which the first bubble of vapor is formed; this temperature is the “bubble point” (see Fig. 1c). The bubble is richer than the liquid in the more volatile components of the mixture. As heating continues, more vapor is formed *and* the temperature continues to

rise, though not as rapidly as before; that is, both sensible and latent heat transfer are occurring to the process fluid. Thermodynamic phase equilibrium calculations are required to find the amount and composition of the vapor phase, the temperature of the fluid, and the amounts of sensible and latent heat transfer. These calculations are an essential part of the design of any heat exchanger involving phase changes of multicomponent mixtures. If heating is continued and the liquid and vapor phases are kept in intimate contact, the last liquid (rich in the less volatile components) vaporizes at the “dewpoint” temperature. Further heating results in superheating the vapor.

Another common problem is the condensation of vapor from a distillation column, possibly using water or air as the coolant. The vapor may be either a nearly pure chemical species or a multicomponent mixture. If nearly pure, the vapor will condense almost isothermally at its saturation temperature corresponding to the vapor pressure, as shown in Fig. 1d. If multicomponent, the vapor will begin to condense at its dew point and continue through the two-phase region until it reaches the bubble point and is totally condensed, as in Fig. 1e. Through the two-phase region, both condensation (latent heat transfer) and cooling of the mixture (sensible heat transfer) occur simultaneously. If the condensate is further cooled below the bubble point, the liquid is said to be subcooled.

The above examples have the common feature of the thermal condition of the process fluid being altered by the use of steam for heating or air or water for cooling. Steam (usually available at several pressures), water, and air are often termed “service” or “utility” streams, and have the common feature of being generally supplied throughout the plant as required. Other service streams include special high-temperature heat transfer liquids such as Dowtherm, hot oil, and occasionally liquid metals; sea water and various refrigerants may also be available as coolants.

Use of service streams to thermally modify process streams is simple, convenient, and operationally flexible, but it is inefficient in terms of energy conservation. Steam has to be made by burning a fuel; cooling water has to be cooled in a cooling tower. In the typical process plant, there are many hot streams that need to be cooled and many cold streams that need to be heated. If the temperatures, flow rates, and locations within the plant are satisfactory, a hot process stream can be used to heat a cold process stream in a heat exchanger (which in this case is often called a feed-effluent exchanger), resulting in a more energy-efficient plant. As an example, the hot vapor effluent (product) stream from an exothermic chemical reactor may be used to heat the cold feed stream to that reactor. While each stream may pass through all of the processes described above, a more typical situation is one in which the cold feed steam starts out as a two-phase gas/vapor–liquid mixture and is totally vaporized

and then superheated, while the hot effluent stream enters the exchanger as a superheated vapor and is then cooled and partially condensed. This case is diagrammed in Fig. 1f.

It is evident that a wide variety of heat transfer processes occurs in heat exchangers in chemical process plants, and, like snowflakes, no two cases are identical. The task of the engineer is to select and properly size a heat exchanger, or a system of heat exchangers, to accomplish the desired thermal changes in the process streams.

II. CRITERIA FOR SELECTION

Given the large variety of process heat transfer problems and the heat exchanger configurations available, the engineer must select a type and design that satisfy several criteria. These are listed approximately in the order of their importance, though in any individual case one criterion or another may move up or down in the list of relative importance.

1. The heat exchanger must satisfy process specifications; that is, it must perform the required thermal change on the process stream within the pressure drop limitations imposed. The basic thermal design equations are discussed in a later section, and these determine the size of the heat exchanger. Equally important to a successful design is the proper utilization of the allowed pressure drops for each stream. As a general rule, the greater the allowable pressure drop, the higher the fluid velocity and heat transfer coefficient, resulting in a smaller and less expensive heat exchanger. However, pressure drop increases with fluid velocity more rapidly than does heat transfer, and pumping costs soon become prohibitive. Also, excessive velocities can cause damage by cavitation, erosion, and vibration. Therefore, the allowable pressure drop in each stream should be carefully chosen (70 kPa is a typical value for low-viscosity liquids, and 5–10% of the absolute pressure is typical for low-pressure gases and vapors), and as fully utilized as other considerations permit.

2. The heat exchanger must withstand service conditions. The most obvious condition is that the exchanger construction must be strong enough to contain the fluid pressures inside the exchanger, and design standards for safe construction are set by the various pressure-vessel codes. There are also thermally induced stresses due to the differential expansion of the various exchanger components. There are mechanical stresses imposed by the exchanger weight and externally by piping stresses, wind loading, and mechanical handling during shipping, installation, and maintenance. The heat exchanger must withstand corrosive attack, primarily achieved by suitable selection of the materials of construction. To minimize

erosion and vibration problems, it is important to limit velocities, especially in certain critical areas near the nozzles and wherever the flow is forced to change direction in the heat exchanger. The exchanger must also be designed either to minimize fouling or to withstand the mechanical effects as fouling does develop.

3. The heat exchanger must be maintainable. It must allow mechanical or chemical cleaning if the heat transfer surface becomes fouled, and it must permit replacement of the tubes, gaskets, and any other components that may fail or deteriorate during the normal lifetime of the exchanger. Maintenance should be accomplished with minimum downtime and handling difficulties and labor cost.

4. Operational flexibility. The heat exchanger and its associated piping and control system must permit operation over the probable range of conditions without instability, excessive fouling, vibration problems, or freeze-up that might damage the exchanger itself. Both changes in process conditions (e.g., changes in process flow rate or composition) and in environmental conditions (e.g., daily and seasonal changes in atmospheric temperature) must be considered.

5. Cost. Cost considerations must include not only delivered cost and installation, but particularly the cost of lost production. The value of products from a process plant is generally so much greater than the cost of any one piece of equipment that loss of production due to inadequate equipment capacity or excessive downtime quickly outweighs any capital cost savings achieved by undersizing equipment.

6. Other design criteria include maximum weight, length, and/or diameter limitations to facilitate installation and maintenance. Use of standard replaceable components minimizes inventory.

III. TYPES OF HEAT EXCHANGERS

Many different types of heat exchangers are available for use in chemical engineering applications, and each has its special features that make it more or less desirable for any given application. A few of the most common types will be described here, together with the advantages, disadvantages, and areas of greatest use.

A. Double-Pipe Exchangers

A typical double-pipe exchanger is shown in Fig. 2. It consists of two concentrically arranged pipes or tubes, with one fluid flowing in the inner pipe and the other in the annulus between the pipes. Special end fittings are used to get the fluids into and out of their respective flow channels and keep them from leaking to the atmosphere. Additional

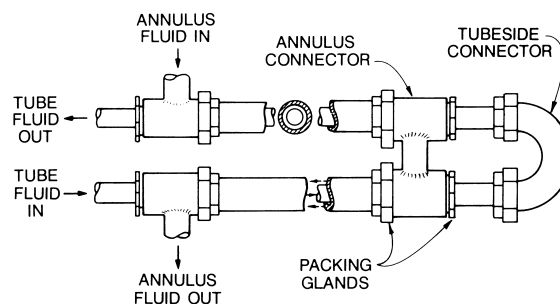


FIGURE 2 Double-pipe heat exchanger.

double-pipe sections can be added in series or parallel to provide the required amount of heat transfer surface.

The double-pipe exchanger is very flexible: vaporization, condensing, or single-phase convection can be carried out in either channel, and the exchanger can be designed for very high pressures or temperatures if required. By proper selection of diameters and flow arrangements, a wide variety of flow rates can be handled. The exchanger can be assembled quickly from standard components and equally quickly expanded or reconfigured if process requirements change. The inner tube can be finned longitudinally on either the internal or the external surface, or both, if additional heat transfer is required in contact with a fluid with poor heat transfer capability. However, the double-pipe exchanger is comparatively heavy, bulky, and expensive per unit of heat transfer area, and it is usually limited to exchangers with less than about 20 m² of surface.

A related design is the multitube, or hairpin, unit, having several internal tubes (usually finned) in a single outer tube, giving a much larger heat transfer area per unit.

B. Shell-and-Tube Heat Exchangers

Shell-and-tube exchangers are the workhorses of the process industries, because they provide a great deal of heat transfer surface in a mechanically rugged configuration and offer so much design flexibility to meet the special requirements of a particular application. Shell-and-tube exchangers are commonly designed to operate at pressures to 200 atm (20 MPa) or temperatures to 650°C, with special designs going higher. Figure 3 is a schematic of a typical shell-and-tube exchanger, showing the principal components, described below.

(A) *Tubes* provide the effective heat transfer area between the fluids, with one fluid flowing inside the tubes and the other fluid flowing across the tubes on the outside. The tubes may be “plain” or “bare,” i.e., having a smooth surface, or they may be “finned,” having from 400 to 1600 fins/m. These fins are radial, like a pipe thread,

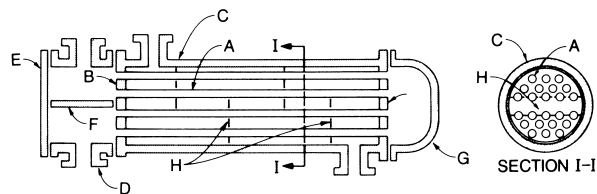


FIGURE 3 Sectional view of a typical fixed tubesheet shell-and-tube heat exchanger: (A) tubes; (B) tubesheets; (C) shell; (D) tube-side channel and nozzles; (E) channel cover; (F) pass divider plate; (G) stationary rear head (bonnet type); (H) tube support plates (or baffles).

and about 0.6–1 mm high and about 0.2–0.4 mm thick. The fins give from $2\frac{1}{2}$ to 5 times the outer surface area of a plain tube. The tubes are arranged in a repeated geometric pattern, usually square or triangular, with the distance between centers of the tubes 1.25–1.5 times the tube outside diameter. The tubes are held in place at each end by being inserted into holes drilled in the tubesheets (B) and welded or roller-expanded into grooves.

(B) *Tubesheets* hold the tubes in place and provide the barrier between the tube-side fluid in the tube-side channels or head and the shell-side fluid. The tubesheet is a circular plate, thick enough to withstand any pressure difference between the two fluids and suitably drilled to accept the tubes. The tubesheets may be welded to the shell (C) as in the diagram, or bolted to a flanged shell, giving a “fixed tubesheet” design. “Floating head” designs are described below. The tubes are fastened to the tubesheets either by welding or by cutting two circumferential grooves in the tubesheet around each tube hole and expanding the tube into the grooves after it is inserted into the tube hole. The expanded tube joint is as strong as a welded joint, but welding is more effective in preventing leaks.

(C) The *shell* confines the flow of the shell-side fluid; the arrangement of the nozzles defines the flow path relative to the tubes. The more common shell and shell nozzle arrangements are shown with their Tubular Exchanger Manufacturers Association (TEMA) designations in Fig. 4 (middle column). The E shell, with the inlet and outlet nozzles at opposite ends of the shell, is the most common arrangement. The K shell is most commonly used when a liquid on the shell side is to be boiled; the large-diameter shell above the tube bundle provides a disengaging area in which most of the droplets of liquid can separate from the vapor leaving through the upper nozzle. The other shell configurations are used to meet specialized requirements in low pressure drop, improved thermal efficiency, or boiling or condensing applications. The shell may be constructed, especially in the smaller diameters, from a length of steel pipe, or it may be rolled from a steel plate and welded along the abutting edges. Holes are cut in the shell at the desired points and the nozzles are welded in.

(D) Tube-side *channels* and *nozzles* control the flow of the tube-side fluid into and out of the tubes. The tube-side channel may be bolted to the shell by flanges (as shown in the drawing) or welded directly.

(E) The *channel cover* bolts over the end of the channel and contains the tube-side fluid. It may be easily removed to allow inspection, cleaning, or replacement of the tubes without disturbing the tube-side piping.

(F) A *pass divider*, or *partition plate*, is used in the case illustrated in order to cause the tube-side fluid to flow through just half of the tubes before turning around in the bonnet (G) to flow back through the other half of the tubes. This results in a *two-pass* configuration, with the liquid flowing through the tubes at twice the velocity that it would have otherwise and allowing the outlet tube-side nozzle to be at the same end of the exchanger as the inlet. More complex pass divider arrangements permit four, six, etc., tube-side passes. The higher tube-side velocities improve the heat transfer rate and tend to minimize fouling (dirt accumulation) on the surface, but the increasing pressure drop and erosion put a limit to the number of passes that can be used. The pass dividers have to be gasketed against the tubesheet and the channel cover to prevent leakage of fluid directly from the inlet to the outlet without passing through the tubes.

(G) The *bonnet* shown here confines the tube-side fluid exiting from the first-pass tubes and turns it around into the second-pass tubes. The bonnet and the channel are basically interchangeable—a channel (with no pass divider) could have been used here instead of the bonnet, or a bonnet with welded-on nozzles could have been used instead of the channel at the inlet/exit end. The channel/channel cover combination is more expensive and more prone to leakage, but allows tube inspection, etc., without disturbing the piping connections.

(H) The *baffles* are required to support the tubes against vibration and sagging, and also to guide the shell-side fluid into a predominantly cross-flow pattern across the tube bundle. The baffles are usually circular plates with an outside diameter slightly less than the shell inside diameter and cut segmentally to provide a window for the fluid to flow from one cross-flow section to the next. Holes are drilled in the baffles for the tubes to pass through, the diameter of the tube holes being slightly greater than the outside diameter of the tubes. The baffle cut (i.e., the height of the segment cut from the baffle) varies from about 15% to 25% of the shell inside diameter for liquids to about 45% for low-pressure gases. Other baffle/tube support geometries are used for special purposes.

Because the shell and the tubes in a heat exchanger are at different temperatures, they will expand by different amounts. The resulting thermal stresses can easily be high enough that tubes are pulled out of tubesheets, or pulled

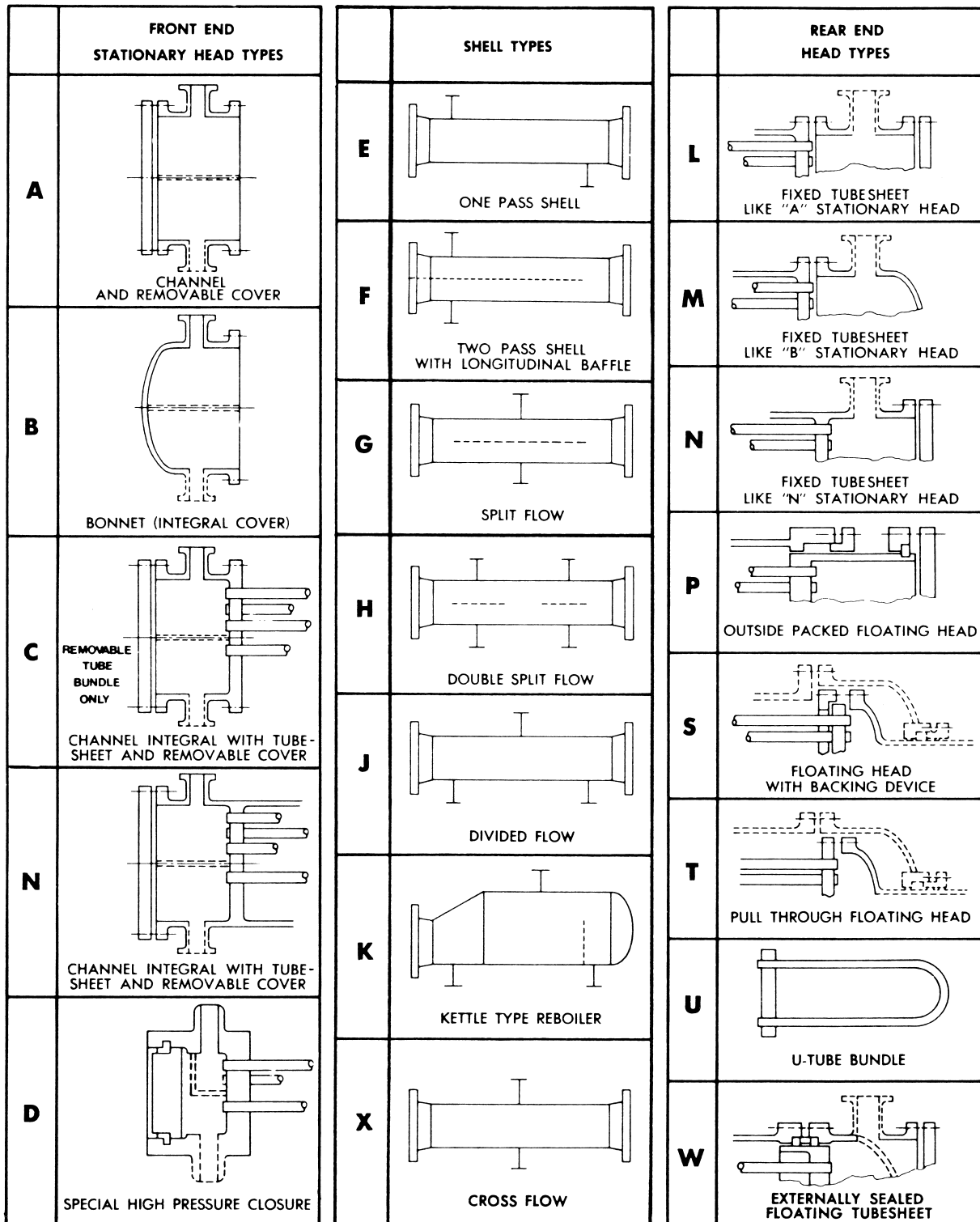


FIGURE 4 Standard notation system for major types of shell-and-tube heat exchangers. [From "Tubular Exchanger Manufacturers Association Standards." (1988). 7th ed. TEMA, New York. ©1985 by Tubular Exchange Manufacturers Association.]

apart or bowed, or the shells badly distorted. The fixed-tubesheet exchanger shown in Fig. 3 can only be used with small temperature differences (typically less than 50°C between the entering fluid streams). Other configurations must be used when thermal stress is a problem.

The U-tube exchanger (Fig. 4) is the best solution to the thermal stress problem, because each tube is free to expand or contract independently of the others or of the shell. However, there are disadvantages to the U-tube exchanger: the tubes cannot be mechanically cleaned around the bend, the inner tubes cannot be individually replaced, the long-radius U-tubes are particularly subject to vibration, and single-pass tube-side flow is not possible.

Therefore, several different designs of “floating-head” exchangers (Fig. 4, rear-end head types P, S, T, and W) are commonly used to relieve the thermal stress problem; the characteristic feature is that the assembly of the rear tubesheet and the associated head is not mechanically connected to the shell. The choice among the configurations depends on the pressures, the temperatures, and the degree of danger should leakage occur to the atmosphere or between the two fluids.

To reemphasize, shell-and-tube heat exchangers are the most commonly employed type in the chemical process industries because they are so adaptable to such wide ranges of conditions.

C. Gasketed-Plate Heat Exchangers

A gasketed-plate heat exchanger is illustrated in Fig. 5, and representative plates are shown in Fig. 6. In this type of exchanger, the heat transfer surface is the thin corrugated

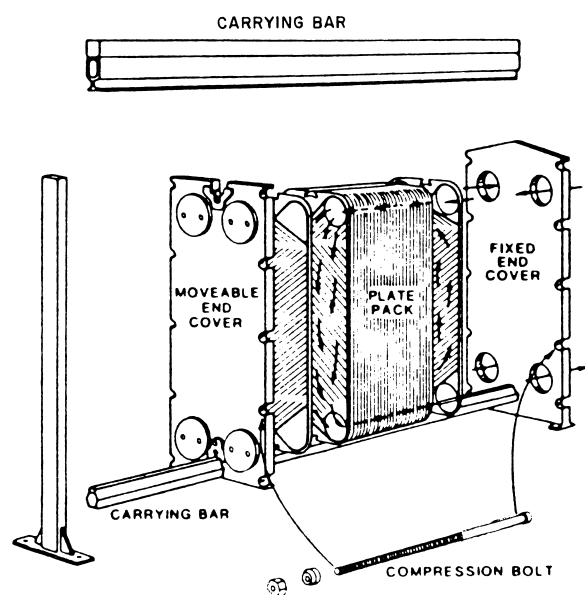


FIGURE 5 Exploded view of a gasketed-plate heat exchanger.

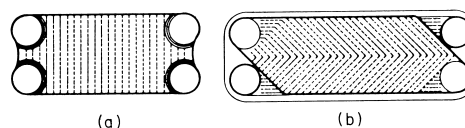


FIGURE 6 Two types of gasketed plates for the gasketed-plate heat exchanger: (a) parallel-corrugated plate; (b) cross-corrugated plate (“herringbone” pattern or “chevron”).

metal plate separating the two fluids. Each fluid flows between adjacent pairs of plates, the gasketing arrangements around the corner ports determining which fluid will flow between each pair, and the gasket around the outer edge sealing each fluid against leakage to the atmosphere. The individual plates are corrugated so that they will mutually support each other against pressure differences between the two fluids. The corrugations cause the flow to be very turbulent, resulting in high heat transfer coefficients and high pressure drops. The strong turbulence also tends to minimize fouling on the surface.

The stack of plates is pressed together by the compression bolts to seat the gaskets. The plates may be made of any metal that can be pressed and provide more heat transfer surface per unit mass of metal and per unit volume than a shell-and-tube exchanger. Therefore, gasketed plate exchangers cost much less per unit surface area than shell-and-tube exchangers if a metal other than low-carbon steel must be used. However, the gasket (usually made of a synthetic rubber or polymer) limits plate heat exchangers to pressures below 20 atm (2 MPa) and temperatures below 175°C, with lower limits for the larger sizes.

In the smaller sizes, the gasketed plate exchanger can be easily taken apart for cleaning and sterilization, so they are widely used in the food processing industry. The larger sizes are used in chemical processing where stainless steel or other high alloys are required. The largest sizes (up to 2200 m² in a single unit) are often constructed of titanium and used with sea water on one side as a cooling medium.

D. Plate-Fin Exchanger

A cutaway view of a plate-fin heat exchanger (sometimes called a matrix exchanger) is shown in Fig. 7. This type of exchanger is built up with alternate layers of matrix sheets (usually corrugated aluminum) and parting sheets. Different fluids flow through alternate layers of matrix, separated from one another by the parting sheets. Sealing against the outside is accomplished by solid side bars, and each fluid is distributed to its respective layers of matrix by a header system. By appropriate header design and location, the plate-fin exchanger can handle three or more separate streams in the same exchanger. After the layers are built up to the desired configuration, the assembly is permanently brazed together using a

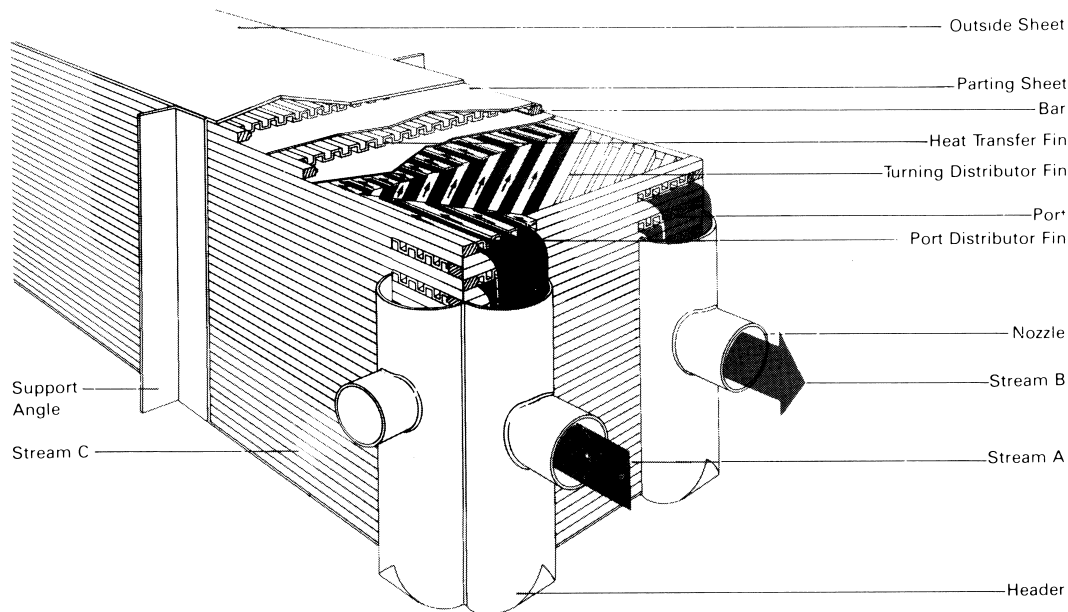


FIGURE 7 Cutaway view of a brazed aluminum plate fin (or matrix) heat exchanger. [Courtesy ALTEC International, Inc., La Crosse, Wis; formerly the Trane Company.]

salt bath or furnace. The usual material of construction is aluminum.

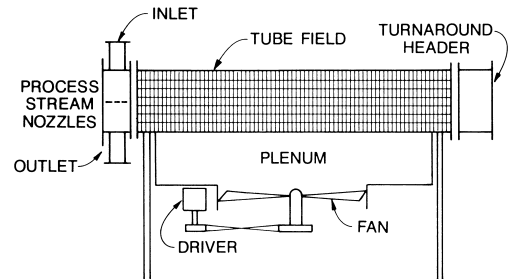
Plate-fin exchangers provide a very large heat transfer surface per unit volume and are relatively inexpensive per unit area. They are not mechanically cleanable and are ordinarily used only with very clean fluids. This combination of properties fits them very well for a wide variety of cryogenic applications, such as air separation; helium separation, purification, and liquefaction; liquefied natural gas production; and separation of light hydrocarbons. They are also used in higher-temperature gas-to-gas services.

E. Air-Cooled Exchangers

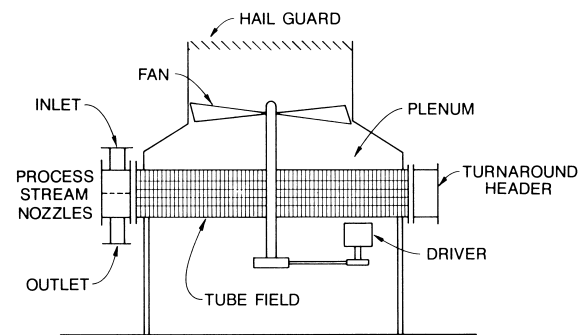
As the name implies, air-cooled exchangers are especially designed to use air as the cooling medium to dissipate low-temperature waste heat. This has become increasingly important in recent years as sources of cooling water have become scarcer and subject to environmental controls. The two basic designs of air-cooled exchangers are shown in Fig. 8. In the forced-draft design, the air is blown upward across the tube field (the heat transfer surface proper) by the fan; in the induced-draft configuration, the air is drawn across the tube field. Units operating at higher exit air temperatures will likely be forced draft to keep the fan out of the hot air; units operating close to ambient air temperature will likely be induced draft so that the plume of warm exhaust air will be more strongly dispersed into the atmosphere, minimizing the possibilities of recirculation.

The critical factor determining the configuration of air-cooled exchangers is the low density and poor thermal

conductivity and specific heat of the cooling medium, air. The low density and specific heat mean that very great volumes of air must be moved through the exchanger to remove the heat from the process fluid. The single-stage



(a)



(b)

FIGURE 8 Typical air-cooled heat exchangers: (a) forced draft; (b) induced draft.

axial-flow fan is the most effective way to move these volumes of air, but it is only capable of very small pressure rises, on the order of 150–500 Pa, with the lower figure being a common design value. The low allowable pressure drop means that the air must be moved very slowly (around 3 m/s) and that the tube field must be shallow (3–12 rows of tubes) and very broad to accommodate the high volumetric flow requirements. This in turn results in low heat transfer coefficients on the air side (about 60 W/m² K).

As will be seen in a later section, it is desirable to provide “extended surface” or fins on an outer tube surface that is in contact with a fluid having a low heat transfer coefficient when the fluid inside the tube has a much higher coefficient. This is almost always the case with air-cooled exchangers, so the tubes used in these exchangers have 350–400 radial fins per meter of length, each fin being about 15–18 mm high. The fins are usually aluminum, about 0.4 mm thick, and wrapped continuously on the tube circumference under tension with a small L-foot to ensure good thermal contact. (Other metals, dimensions, and means of attachment are available commercially.) The result is an effective outside area of the tube about 20 times the inside area.

Air-cooled exchangers require large plan areas, and adequate provision must be made for cool air to flow into the underside of these units. Installations covering 10⁴–10⁵ m² of land are becoming more common.

F. Mechanically Aided Heat Exchangers

Some heat transfer problems require the use of locally applied mechanical energy to achieve acceptable heat-transfer rates. Two typical cases of this type are shown in Figs. 9 and 10.

The first case is typified by a stirred-tank chemical reactor in which heat must be externally added to or removed

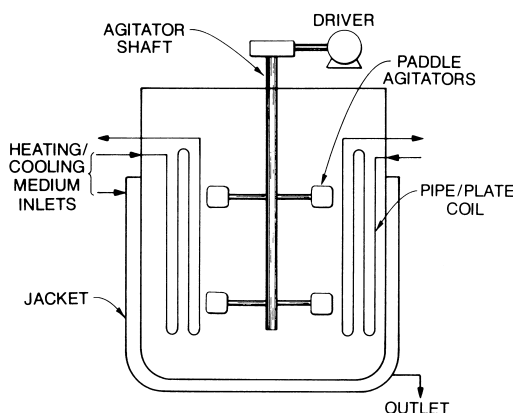


FIGURE 9 Sectional view of a stirred-tank reactor/heat exchanger with both an external jacket and internal heat transfer coils.

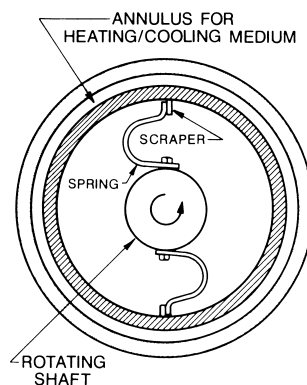


FIGURE 10 Sectional view of a close-clearance, mechanically aided heat exchanger, with spring-loaded blades.

from the chemical reaction in order to control it. Sometimes it is necessary to add heat to start the reaction and then remove heat in order to keep it under control. The heat transfer surface may be either external to the reactor volume (the jacket) or internal (the pipe/plate coils). The mechanical agitation may be provided by paddles (shown), propellers, turbines, helical flights, or jets, or combinations of these. The reactor design is likely to be controlled by the chemical reaction kinetics, but the heat transfer surface and the cooling/heating medium must be sufficient to provide control.

The scraped surface, or close-clearance exchanger, illustrated in Fig. 10 is required for a few very difficult situations. An example is purification by fractional crystallization, in which a refrigerant boiling in the annulus cools a solution of various substances and certain species selectively crystallize out on the surface of the inner pipe. The crystalline deposit must be continuously scraped off of the surface in order that the heat transfer rate be maintained. The crystals are eventually removed from the remaining liquid by filtration.

Mechanically aided heat exchangers are expensive, use large power inputs, and require frequent maintenance. Nonetheless, they are often the only way to accomplish certain tasks.

G. Other Types of Heat Exchangers

The above descriptions cover the major types of heat exchangers in the process industries, but there is a very long list of other configurations that have vital if limited applications. Briefly, these include tube bundles made of Teflon because of its great resistance to chemical attack, spiral plate exchangers with a high area-to-volume ratio and a particular resistance to fouling, welded-plate exchangers and heavy-duty welded-fin exchangers for high-temperature heat recovery, and graphite block exchangers

for resistance to chemical attack and high thermal conductivity.

IV. BASIC HEAT EXCHANGER EQUATIONS

A. Heat Balance

The heat required to heat a fluid that does not change phase from t_i is t_o is

$$\dot{Q}_T = \dot{m}c_p(t_o - t_i), \quad (1)$$

where $\dot{Q}_T$ is the sensible transfer rate (watts or joules per second); $\dot{m}$ is the mass flow rate of the fluid (in kilograms per second), c_p is the mean specific heat of the fluid over the temperature range (in joules per kilogram per kelvin), and t_i and t_o are the inlet and outlet temperatures (in kelvins) of the fluid, respectively. The corresponding heat given up by the hot fluid, assuming it does not change phase and that there are no heat leaks, is

$$\dot{Q}_T = \dot{M}C_p(T_i - T_o), \quad (2)$$

where the terms have similar meanings, applied to the hot fluid.

If, rather, the hot fluid is an isothermally condensing vapor (such as steam), the latent heat duty is

$$\dot{Q}_T = \dot{M}\lambda, \quad (3)$$

where $\dot{M}$ is the mass rate of condensing (in kilograms per second) and λ is the latent heat of condensation (in joules per kilogram) at the condensing temperature.

If a fluid composed of more than one component (e.g., a solution of ethanol and water, or a crude oil) partially or totally changes phase, the required heat is a combination of sensible and latent heat and must be calculated using more complex thermodynamic relationships, including vapor-liquid equilibrium calculations that reflect the changing compositions as well as mass fractions of the two phases.

B. Rate Equation

Consider the typical case of heat transfer between one fluid inside a tube and another fluid outside the tube, shown in cross section in Fig. 11. Heat is transferred by convection from the hot fluid (taken arbitrarily to be the fluid inside the tube) to the fouling deposit (if any) on the inside surface, through the fouling deposits and tube wall by conduction, and then by convection to the fluid outside the tube. At the point where the inside fluid temperature is T and the outside is t , the local heat flux inside the tube is

$$\dot{q}_i = \frac{d\dot{Q}}{dA_i} = U_i(T - t), \quad (4)$$

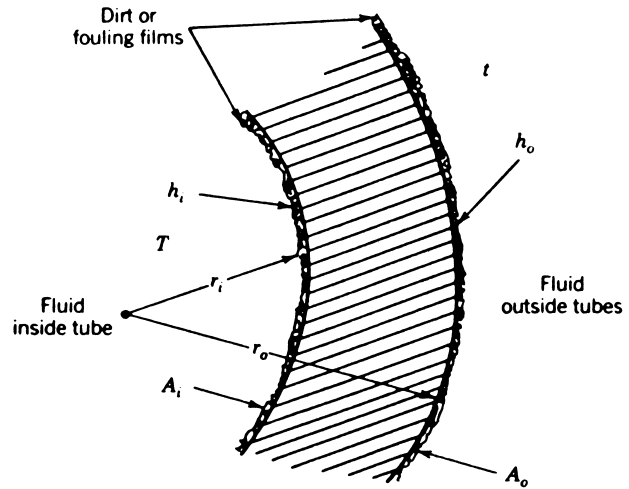


FIGURE 11 Cross section of a heat exchanger tube, with convective heat transfer in the fluids and fouling deposits on the surfaces.

where $\dot{q}_i$ is the local heat flux (in watts per square meter or joules per square meter per second), $d\dot{Q}$ is the differential amount of heat transferred through the differential heat transfer area (inside surface area) dA_i (in square meters), U_i is the overall heat transfer coefficient based on the inside heat transfer area (in watts per square meter per kelvin or joules per second per square meter per kelvin), and T and t are the local hot and cold fluid temperatures (in kelvins).

The overall heat transfer coefficient is related to the individual heat-transfer processes by the equation

$$U_i = \frac{1}{(1/h_i) + R_{f_i} + (r_i/k_w) \ln(r_o/r_i) + (1/h_o + R_{f_o}) \left(\frac{A_i}{A_o}\right)}, \quad (5)$$

where h_i and h_o are the convective heat transfer coefficients (in watts per square meter per kelvin or joules per second per square meter per kelvin) for the inside and outside fluids, respectively, each based on the corresponding area. A_i and A_o ; R_{f_i} and R_{f_o} are the inside and outside fouling resistances (in square meters-kelvins per watt or second-square meters-kelvins per joule), each based on the corresponding area; r_o and r_i are the inside and outside radii of the tube, k_w is the thermal conductivity of the tube wall (watts per meter per kelvin or joules per second per meter per kelvin), and A_i and A_o are the inside and outside surface areas of the tube (in square meters). Strictly speaking, the above equation applies only to plain cylindrical tubes for which

$$A_i = 2\pi r_i L, \quad (6a)$$

$$A_o = 2\pi r_o L, \quad (6b)$$

where L is the tube length. However, the equation can be applied with small modifications to tubes with external fins, where A_o now is the total heat transfer surface on the outside of the tube, including the fins. Corresponding to Eq. (5), the overall heat transfer coefficient could have been based on the outside area of the heat transfer surface A_o :

$$U_o = \frac{1}{(1/h_i + R_{fi})(A_o/A_i) + (r_o/k_w)\ln(r_o/r_i) + \frac{1}{h_o} + R_{fo}}. \quad (7)$$

Note that $U_i A_i = U_o A_o$.

The convective heat transfer coefficients h_i and h_o must be calculated from equations that involve the geometry of the system, the physical properties of the fluid, and the velocity with which it is flowing. These equations are obtained variously by more or less fundamental analysis of the heat transfer and fluid flow mechanisms, or by correlation of experimental data, or by combinations of these methods. A few typical values of the film coefficients are

Air, atmospheric pressure, flowing at a few meters per second, 50–100 W/m² K

Water, 1–2 m/s, 4000–6000 W/m² K

Gasoline, 1–2 m/s, 1000–1500 W/m² K

Liquid sodium, 25,000–30,000 W/m² K

Condensing steam, atmospheric pressure, 8,000–15,000 W/m² K

Boiling water, atmospheric pressure, 15,000–25,000 W/m² K

C. The Design Integral and the Mean Temperature Difference

Equation (4) applies at a point in a heat exchanger where the hot and cold fluid temperatures are T and t , respectively. Since one or both of these temperatures will almost always change from point to point in the heat exchanger, depending on the amount of heat exchanged and the flow paths of the two fluids, Eq. (4) must be integrated over the total heat duty of the heat exchanger $\dot{Q}_T$, with T , t , and possibly U_i being expressed as functions of $\dot{Q}$; the integration may be formally expressed as

$$(A_i)_T = \int_0^{\dot{Q}_T} \frac{d\dot{Q}}{U_i(T-t)}, \quad (8a)$$

where $(A_i)_T$ is the total heat transfer area in the heat exchanger (based on the inside area of the tubes) required to transfer $\dot{Q}_T$ (watts or joules per second). Alternatively, the total *outside* surface area required is

$$(A_o)_T = \int_0^{\dot{Q}_T} \frac{d\dot{Q}}{U_o(T-t)}. \quad (8b)$$

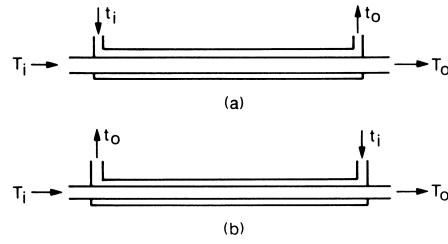


FIGURE 12 Two possible flow arrangements in a heat exchanger: (a) concurrent; (b) countercurrent.

Equations (8a) and (8b) can be analytically integrated if certain assumptions are valid. Key among these assumptions are that the specific heats of each fluid are constant (or that one or both fluids are changing phase isothermally), that the overall heat transfer coefficient is constant throughout the heat exchanger, and that the flows are either entirely cocurrent or entirely countercurrent to one another through the heat exchanger, as illustrated in Fig. 12. The integrations result in

$$(A_i)_T = \frac{\dot{Q}_T}{U_i(\text{LMTD})}, \quad (A_o)_T = \frac{\dot{Q}_T}{U_o(\text{LMTD})}, \quad (9)$$

where LMTD, the logarithmic mean temperature difference, is

$$\text{LMTD} = \frac{(T_i - t_i) - (T_o - t_o)}{\ln[(T_i - t_i)/(T_o - t_o)]} \quad (10a)$$

for cocurrent flow and

$$\text{LMTD} = \frac{(T_i - t_o) - (T_o - t_i)}{\ln[(T_i - t_o)/(T_o - t_i)]} \quad (10b)$$

for countercurrent flow.

If the flows are not entirely cocurrent or entirely countercurrent (as in multipass shell-and-tube exchangers, or in air-cooled exchangers) but the other assumptions are satisfied, Eq. (9) can usually be put in the form

$$(A_i)_T = \frac{\dot{Q}_T}{U_i F(\text{LMTD})_{cc}}, \quad (11)$$

$$(A_o)_T = \frac{\dot{Q}_T}{U_o F(\text{LMTD})_{cc}},$$

where $(\text{LMTD})_{cc}$ refers to the logarithmic mean temperature difference for countercurrent flow, Eq. (10b), and F is an analytically obtained correction factor ($F \leq 1.00$) that is a function of the terminal temperatures of the two streams. Treatment of F calculations is beyond the scope of this article. Many heat exchangers can be and are satisfactorily designed by hand calculations using Eqs. (5) or (7), (10b), and (11), but most exchangers are designed using computer programs based on the numerical integration of Eq. (8a) or (8b).

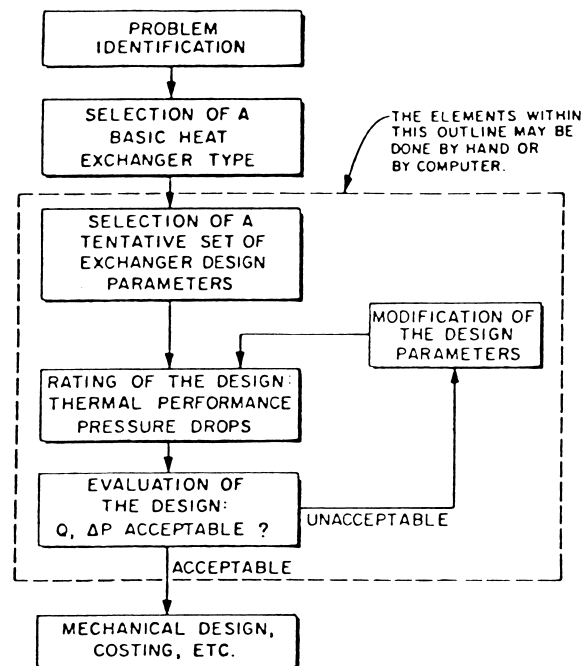


FIGURE 13 Structure of the heat exchanger design process.

V. THE DESIGN PROCESS

The structure of the process heat exchanger design procedure is shown in Fig. 13. The basic structure is the same whether hand or computer-based design methods are used; all that is different is the replacement of the very subtle and complicated human thought process by an algorithm suited to a fast but inflexible computer.

First, the problem must be identified as completely and unambiguously as possible, including stream compositions, flow rates, and temperatures, and the likely ranges of variations in these parameters during operation. Any design problem will have certain contextual considerations the designer needs to know in order to arrive at a near-optimal design. A major judgment, usually made almost instinctively, is the level of engineering effort justified by

the actual value of the exchanger in the process. Also, at this point the single most important decision is made (often by default): what basic configuration of exchanger is to be chosen and designed? In the process industries the usual answer is the shell-and-tube exchanger, but it is always worth reviewing the other possibilities.

The next decision is what design method is to be used. Basically, these fall into two categories: hand design and computer design. Hand design methods in the most recent literature and applied by a competent designer are still valid for a large fraction of all heat exchanger problems. If one chooses to use a computer design method, there is still the task of selecting the level of the method. There are short-cut and detailed computer design methods available for most exchanger types.

The next step is to select a tentative set of exchanger geometric parameters. The better the starting design, the sooner the designer will come to the final design, and this is very important for hand calculation methods. On a computer, however, it is usually faster to give the computer a very conservative (oversized) starting point and use its enormous computational speed to move toward the desired design.

In either case the initial design will be “rated”; that is, the thermal performance and the pressure drops for both streams will be calculated for this design. The rating program is diagrammed in Fig. 14. In the rating program the problem specifications and the preliminary estimate of the exchanger configuration are used as input data; the exchanger configuration given is tested for its ability to effect the required temperature changes on the streams within the pressure-drop limitations specified.

The rating process carries out three kinds of calculations. It first computes a number of internal geometry parameters that are required as further input into the heat transfer and pressure drop correlations. Then the heat transfer coefficient and pressure-drop are calculated for each stream in the configuration specified.

The results from the rating program are either the outlet temperatures of streams, if the length of the heat exchanger has been fixed, or the length of the heat exchanger required

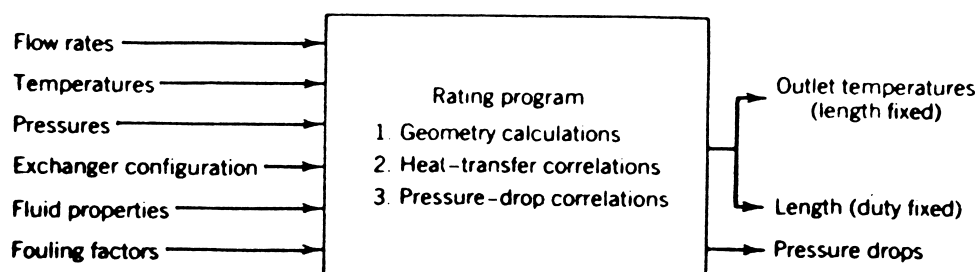


FIGURE 14 The rating program.

to effect the necessary thermal change if the duty has been fixed. In either case the rating program will also calculate the pressure drops for both streams in the exchanger.

If the calculation shows that the required amount of heat cannot be transferred or if one or both allowable pressure drops are exceeded, it is necessary to select a different, usually larger, heat exchanger and rerate. Alternatively, if one or both pressure drops are much smaller than allowable, a better selection of parameters may result in a smaller and less costly heat exchanger, while utilizing more of the available pressure drop.

The design modification program takes the output from the rating program and modifies the configuration in such a way that the new configuration will do a "better" job of solving the heat transfer problem.

A computer-based configuration modification program is a complex one logically because it must determine what limits the performance of the heat exchanger and what can be done to remove that limitation without adversely affecting either the cost of the exchanger or the operational characteristics of the exchanger which are satisfactory. If, for example, it finds that the heat exchanger is limited by the amount of heat that it can transfer, the program will try either to increase the heat transfer coefficient or to increase the area of the heat exchanger, depending on whether or not pressure drop is available. To increase the tube-side coefficient, one can increase the number of tube passes, thereby increasing the tube-side velocity. If shell-side heat transfer is limiting, one can try decreasing baffle spacing or decreasing the baffle cut. To increase area, one can increase the length of the exchanger, or increase the shell diameter, or go to multiple shells in series or in parallel.

Clearly the possibilities are enormous, and the configuration modification program must be very tightly written to avoid wandering off into impossible designs or loops without an exit. A designer using a hand method makes many decisions intuitively, based on the experience the designer has built up. In any case, once a final design has been arrived at by the computer, the basic rationality and approximate correctness of the solution should be verified by an experienced designer.

VI. FURTHER DEVELOPMENTS IN DESIGN AND APPLICATION

There is continuing rapid development in both the hardware and the software of the process heat exchanger industry. New types of heat exchangers appear from time to time, usually arrived at for solving a particular problem that is not quite properly or economically satisfied by existing equipment. Even in well-known types, growth

in the fundamental understanding of the details of the heat transfer and fluid mechanics processes leads to modifications in the structure of the heat exchanger. Particularly, there is now a concerted effort to understand and categorize the fundamental fluid mechanical processes operating in various kinds of "enhanced" surfaces. For example, it has been found that a spirally fluted tube swirls the fluid as it flows through the tube, giving rise to secondary flows in the immediate vicinity of the surface which increase the heat transfer coefficient with little increase in pressure drop. Similar efforts are underway with flows outside tubes and for two-phase (vaporizing or condensing) flows.

There is continuing development of new manufacturing techniques, including explosive bonding of metal parts (e.g., tubes to tubesheets), oven brazing for compact heat exchangers (including stainless steel and titanium), and ceramic heat exchangers for very high temperature applications.

In the software area, computer programs for the analysis design of heat exchangers are moving constantly toward using more fundamental and detailed calculations of the actual flow field and the thermal transport within the flow field. The improvement in computer capabilities has meant that the design engineer has direct access to the most highly advanced design methods at his desk rather than having to access a mainframe computer in batch mode. Meanwhile, the growing supercomputer availability is allowing the research engineer to study the fundamental fluid mechanical and corresponding thermal transport processes in the complex geometries characteristic of actual heat exchangers.

Heat exchangers are a prime means of conserving energy in process plants by exchanging heat between process streams that need to be cooled and those that need to be heated. Much attention is directed toward optimization of heat conservation and recovery by selecting the proper heat exchanger network.

With all the advances that have been made, there is still room for much more. Our knowledge of the mechanisms of fouling is very limited, and meaningful predictive methods are almost nonexistent. Many important heat, mass, and momentum transfer processes are still poorly understood. Finally, some cases where the basic equations are known still must be crudely approximated in design because the computational requirements for complete design still exceed those of the most powerful computers.

SEE ALSO THE FOLLOWING ARTICLES

DISTILLATION • FLUID MIXING • HEAT TRANSFER • STEAM TABLES • THERMODYNAMICS

BIBLIOGRAPHY

- Hewitt, G. F., ed. (1998). "Heat Exchanger Design Handbook—1998," Begell House, New York.
- Hewitt, G. F., Shires, G. L., and Bott, T. R. (1994). "Process Heat Transfer," CRC Press, Boca Raton, FL/Begell House, New York.
- Kakac, S., and Liu, H. (1998). "Heat Exchangers: Selection, Rating, and Thermal Design," CRC Press, Boca Raton, FL.
- Kays, W. M., and London, A. L. (1984), "Compact Heat Exchangers," 3rd ed., McGraw-Hill, New York.
- Saunders, E. A. D. (1988). "Heat Exchangers: Selection, Design and Construction," Longman/Wiley, New York.
- Smith, R. A. (1986). "Vaporisers: Selection, Design and Operation," Longman/Wiley, New York.
- "Standards of the Tubular Manufacturers Association," 7th ed. (1988). Tubular Exchanger Manufacturers Association, Tarrytown, NY.
- Webb, R. L. (1994). "Principles of Enhanced Heat Transfer," Wiley, New York.
- Yokell, S. (1990). "A Working Guide to Shell-and-Tube Heat Exchangers," McGraw-Hill, New York.



High-Pressure Synthesis (Chemistry)

R. H. Wentorf, Jr.
R. C. DeVries

General Electric Corporate Research and Development

- I. General Remarks about Pressure
- II. Pressure as a Thermodynamic Variable
- III. Methods for Generating Very High Pressures
- IV. General Considerations of Phase Changes
- V. Practical Uses of Very High Pressures
- VI. Journey to the Center of the Earth

GLOSSARY

Cubo-octahedron Crystal having the faces of both the cube and the octahedron.

Emf Electromotive force, measured in volts.

Gasket Softer or more deformable plastic material placed between harder materials, usually to seal a gap.

Nucleation Process by which the first tiny groups of atoms or molecules form a nucleus on which a crystal can grow. Growth is easier than nucleation.

Prestressed Part of a structure carries some stress even before the structure is under load. With prestressing, certain structures can carry higher loads because the stresses under load are more uniform.

Stress Force per unit area on any imaginary plane in an object. Stresses perpendicular to the plane are compressive or tensile; stresses parallel with the plane are shear stresses. Up to the elastic limit, stresses produce

no permanent change after they are removed. In the plastic or failure stress range, the body is permanently deformed.

“VERY HIGH PRESSURES” means those generally above 10 kbar (kilobar), or 1 GPa (gigapascal). Confinement methods for almost any substance can usually be found, and working temperatures may range from about 1 to 4000 K. The generation and use of such pressures entail special techniques and small volumes of material; therefore, the industrial uses of such pressures are limited to the preparation of materials of high unit value such as diamond and cubic boron nitride. However, for scientific investigations, small working volumes do not matter so much and very high pressures are widely used, particularly by those interested in the solid state of matter or the interiors of the earth and other planets.

TABLE I Pressure Units and Phenomena

Experience	Pressure			
	Pa	bar	atm	psi
Lift on wing of light plane	690	6.9×10^{-3}	.006805	0.10
Pressure in tire	2.026×10^5	2.026	2	29.4
Cylinder, gasoline engine	10^6	10	9.8692	145
Hydraulic jack, compressed gases	10^7	100	98.692	1450
Ocean depths	10^8	1000	986.92	1.45×10^4
Metal forming	5×10^8	5000	4934.6	7.25×10^4
Diamond synthesis, depths of moon	5×10^9	50000	49,346	7.25×10^5
Center of Earth	3.64×10^{11}	3.64×10^6	3.59×10^6	5.28×10^7
Center of Jupiter	$\sim 10^{13}$	$\sim 10^8$	$\sim 10^8$	$\sim 10^9$
White dwarf star, degenerate matter	$\sim 10^{18}$	$\sim 10^{13}$	$\sim 10^{13}$	$\sim 10^{14}$

I. GENERAL REMARKS ABOUT PRESSURE

Pressure is force per unit area. The modern unit of pressure is the Pascal (Pa), which is a force of 1 N (Newton) on an area of 1 m². A newton accelerates a kilogram at 1 per sec². One pascal represents a pressure which is very small relative to daily experience, as is illustrated in Table I.

Although prehistoric humans used high pressures in the shaping of stone tools and many a medieval blacksmith hammered on cold iron, up until the early 1950s relatively few scientists actively worked in the field of very high pressure. The main exceptions were those who studied high-velocity phenomena in connection with military explosives (shaped charges and atomic bombs), and Professor P. W. Bridgman of Harvard University, who investigated many effects of static pressures up to 10 GPa. He was awarded the Nobel prize in physics in 1946. His pioneering work, described in his books and collected papers, still provides many modern workers with insight and inspiration.

Interest in high-pressure phenomena was reawakened about 1955 by the synthesis of diamond from graphite, and since then many workers from a myriad of disciplines have used very high pressures to explore the behavior of matter. International conferences on high pressures are held every year, and the literature on the subject is large and growing.

II. PRESSURE AS A THERMODYNAMIC VARIABLE

In a solid or liquid at room pressure and temperature, the attractive forces between atoms balance those of thermal agitation and repulsion. As the external pressure on a

substance increases, the interatomic repulsive forces become more noticeable. For most substances a pressure of about 2 GPa makes the repulsive forces predominant, and their stiff nature significantly reduces compressibilities at higher pressures.

Table II sets forth the relative volumes and approximate internal energy changes produced by compression to 10 GPa for a few substances of widely different compressibilities. To compress 1 mm³ of material to 10 GPa is no trivial matter, yet the energy changes seem small compared to heating.

Many more interesting effects of pressure spring from the broadening and overlapping of the outer electronic states of atoms associated with chemical bonding, the distortion of molecules or crystalline arrangements, and the shifts of equilibria associated with volume changes. A volume change of 4 ml g-mol⁻¹ at 10 GPa means a free-energy change of 10 kcal g-mol⁻¹, which is significant compared with the energies of chemical reactions, 20–50 kcal g-mol⁻¹. If, in the course of a chemical reaction, an intermediate state is formed for which the molar volume differs from that of its reacting components, the reaction velocity can be markedly increased or decreased

TABLE II Effects of Compression to 10 GPa at 25°C

Material	Relative volume	Approximate increase in internal energy		
		cal/g	cal/g mol	cal/cm ³
Potassium	0.50	380	15,000	320
Water	0.55	330	6,000	330
NaCl	0.795	60	3,400	130
MgO	0.95	9.4	380	34
Iron	0.95	4.4	240	34
Diamond	0.98	3.5	42	12

by pressure, according to whether the intermediate state is less or more voluminous. The intermediate states in viscous flow or diffusion are more voluminous, and these processes are strongly hindered by pressure.

If the density change on melting is large, pressure will have a large effect on the melting point. The melting temperature of NaCl, for example, rises from 801°C at 1 atm to about 1900°C at 10 GPa. The melting point of iron rises from 1535°C at 1 atm to about 1700°C at 5 MPa. For materials such as bismuth, water, silicon, and probably diamond, the liquid is more dense than the solid and the melting point decreases with pressure. The variation of melting temperature with pressure is given by:

$$dt/dp = \Delta V / \Delta S \quad (1)$$

where ΔS and ΔV are the entropy and volume changes associated with melting.

Generally speaking, it is easy to find a substance that will exhibit some kind of a phase change as the result of compression, but it is more difficult to find a substance that will retain its high-pressure form after the pressure on it is reduced to 1 atm. In most substances the internal bonding of the high-pressure phase is too weak to preserve the structure against decompression or thermal agitation. Hence, most high-pressure forms must be studied at high pressure, and many ingenious devices have been made for such studies. The few high-pressure forms that can be “brought back alive” are typically hard, refractory materials such as carbon, silicon, and silicates. These are usually formed at high temperatures and pressures and then quenched for leisurely study at low pressure.

The problem of recovery leads to the question of hardness. Hard substances have a high number of strongly directed, covalent chemical bonds per unit volume. Soft substances generally have fewer bonds per unit volume or bonds that are weak or weakly directed, such as ionic or dipole attractive forces. Bond energy per unit volume has the same dimensions as pressure (force per unit area), and a plot of hardness measured by the Knoop indenter versus the bond energy per molar volume for various substances is essentially linear, provided that one chooses substances for which the bonding is predominantly of one type (i.e., not mixed, as in graphite or talc).

Covalent (electron pair) bond strengths vary between approximately 60 and 90 kcal/mol for most elements present in hard materials, but the cube of covalent bond length varies even more: approximately 3.65 Å³ for C–C, 6.1 Å³ for Si–O, and 14.3 Å³ for Ni–As. The heavier elements generally offer more bonds per atom, but this usually does not compensate for the larger molar volumes except in certain interstitial compounds such as WC and TiN. Thus, the hardest materials are generally made of

light elements, with diamond at the top. Usually, hard materials are brittle because the strongly directed bonds that favor hardness do not favor plasticity, which involves the intersite motions of atoms during which the attractive forces on the atoms remain relatively constant. However, at sufficiently high ambient pressures many normally brittle materials become plastic as the overall compressive stress makes repulsive forces predominate between atoms. Thus, cracks become energetically unfavorable, although the resistance to deformation may increase. This phenomenon has some applications in industrial processes and in geology.

The long chains of atoms present in oils, greases, and polymers become tightly entangled at high pressures; most atomic displacements then involve breaking of chemical bonds, and the viscosity or shear strength rises markedly. Such “hardening” of oil is probably important in the lubrication of highly stressed areas on cams, gear teeth, etc.

III. METHODS FOR GENERATING VERY HIGH PRESSURES

Two general methods are available. In the “static” method, the substance is confined by the strength of materials and the exposure times are long—seconds to months. In the dynamic method, the substance is confined by inertia and the exposure times are short, of the order of microseconds, due to the difficulty of maintaining large accelerations for long time periods. Nevertheless, the highest pressures are achieved by dynamic methods.

A. Static Apparatus

The simplest apparatus is the piston and cylinder, shown in Fig. 1. The pressure is the force on the piston divided by its area, after allowing for friction and distortion. The strongest practical piston material is cobalt-cemented tungsten carbide. In certain compositions, around 3–6 wt%, cobalt can have a compressive strength of 4–5 GPa along with sufficient ductility to absorb inevitable local high stresses without failure. The strongest cylinders are made with a stiff cemented tungsten carbide inner shell that is supported against bursting (and partly against axial delamination) by prestressed steel rings.

Let us examine the stresses and distortions that accompany the generation of pressure in this apparatus. In Fig. 1 the original (zero pressure) shapes of piston and cylinder are shown by dotted lines; the distortions, shown by the solid lines, due to pressure are exaggerated. The bulging of the piston is most pronounced above the cylinder; inside the cylinder the piston is supported by, and rubs on, the wall of the cylinder. The sharp change in radial bursting

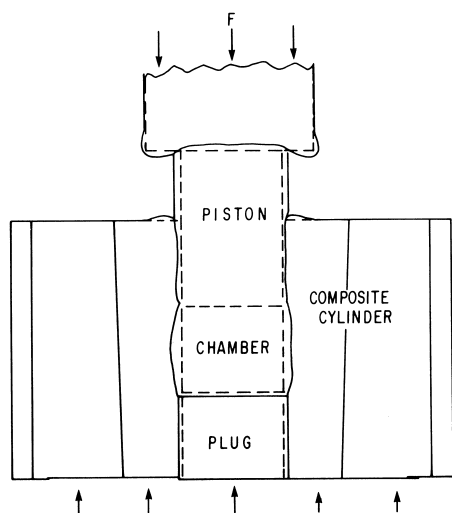


FIGURE 1 Cross section of simple piston and cylinder apparatus at high pressure. Original shape indicated by dashed lines. Distortions due to pressure are exaggerated.

pressure on the cylinder wall at the end of the piston imposes a high shear stress there. This, combined with the pinching effect on the inner cylinder between the high-pressure-chamber contents and the supporting rings, tends to make the inner cylinder separate axially. The situation could be summed up by saying that all the free surfaces tend to bulge and crack.

In order to generate pressures higher than the compressive strength of the piston material, one recognizes that the maximum stress gradients must be kept within limits. This is best done by using a controlled reduction in stress from the inner to the outer parts of the apparatus. This idea is embodied in all very high-pressure apparatus used above about 4 GPa and can be seen in the cross section of the "belt" apparatus shown in Fig. 2. The figure shows the situation at 1 atm and at 6 GPa. A composite conical gasket, consisting of a stone-steel-stone sandwich, seals the gap between piston and cylinder, permits the motion necessary for compressing the chamber contents, provides electrical and thermal insulation, and also supports the piston and cylinder surfaces with a monotonic fall in pressure from the tip of the piston to the atmosphere. Thus, the net stress on the piston falls smoothly, not abruptly, from the tip to the wide base. This point will be discussed more later.

The pressurized volume in this type of apparatus is relatively large, at least a few milliliters, and easily holds an electrically heated furnace. The pistons carry the heating current. Thermocouple or other sensing wires can be led out through the gaskets. Maximum steady temperatures can be as high as 3000°C, depending on the thermal insulation used. Temperatures over 4000°C can be reached in brief (millisecond) pulses. An apparatus of this type

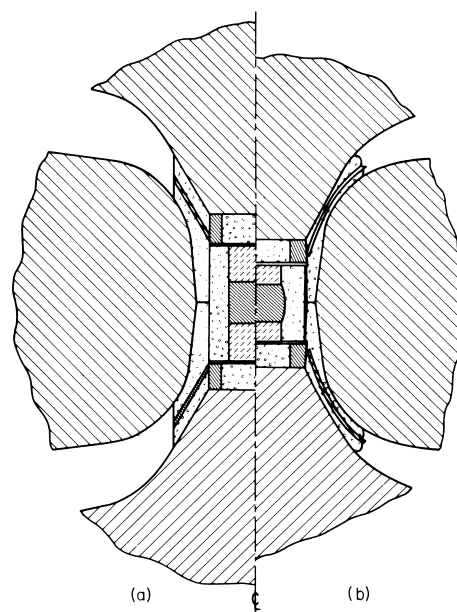


FIGURE 2 Cross section of "belt" high-pressure, high temperature apparatus, split along the center line to show both (a) 1 atm and (b) 6 GPa states.

is suitable for the synthesis of diamond and other high-pressure forms of matter. Special versions can reach pressures of 15 GPa. Sometimes results must be interpreted with care because the chamber pressure can be affected by local density changes resulting from heating or phase transformations. The compressible gasket, though indispensable, makes the determination of chamber pressure more uncertain. Useful pressure calibration methods are discussed later.

Returning now to piston flank support, consider the ideal tapered piston shown in Fig. 3, a truncated cone of half angle α . If h is the slant height measured from the projected apex of the cone, then at $h = h_0$ the working face of the piston is exposed to a chamber pressure

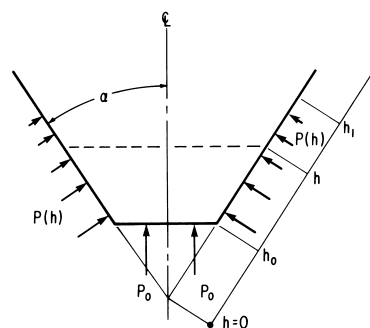


FIGURE 3 Cross section of tapered piston with face pressure P_0 and lateral support pressure $P(h)$.

P_0 on the face area πR_0^2 . Along the flank of the piston, the supporting pressure $P(h)$ falls to zero at $h = h_0$. We require that at any h along the flank, the cross-sectional area of the piston, $\pi h^2 \sin^2 \alpha$, bears a net stress less than the simple compression strength of the material, S . The net stress is the total load on the cross section divided by its area, minus the support pressure $P(h)$. The total load is the piston face load plus the piston flank load.

In symbols this requirement is

$$S + P(h) \geq \frac{1}{\pi h^2 \sin^2 \alpha} \cdot \left(P_0 \pi R_0^2 + 2\pi \sin^2 \alpha \int_{h_0}^h hP dh \right) \quad (2)$$

For the maximum allowable flank pressure gradient the equality holds in Eq. (2).

By differentiating with respect to h we obtain:

$$dP/dh = -2S/h \quad (3)$$

and

$$P = P_0 - 2S \ln(h/h_0) \quad (4)$$

Equation (4) is based on the boundary condition that $P(h_0) = P_0$, the chamber pressure, since it is physically difficult to avoid this situation even though the piston tip is thereby given more support than necessary.

At $h = h_1$, $P = 0$ and we have

$$P_0 = 2S \ln(h_1/h_0) \quad (5)$$

which tells us that in principle any chamber pressure can be confined, but the logarithmic dependence makes it slow going. If we regard h_1/h_0 as a measure of the size of the apparatus, its volume V goes as $(h_1/h_0)^3$, so we have

$$V = B \exp(3P_0/2S). \quad (6)$$

The required piston force F will go as $(h_1/h_0)^2$ so that

$$F = C \exp(P_0/S) \quad (7)$$

where B and C are geometrical constants. The advantages of a high S are obvious, but practical piston materials are limited mostly to cemented tungsten carbide ($S = 5$ GPa) or various forms of diamond ($S \sim 10$ GPa). The latter are available only in small (1 cm) sizes.

Equations (1) to (7) are rather general and apply to conical or pyramidal pistons. For example, if $\alpha = 45^\circ$ and we let $(h_1 - h_0)$ be $6R_0$, the maximum chamber pressure is about $4S$ and approximately 14% of the piston force is due to chamber pressure; 86% is ideally taken by gasket load.

In practice, the ideal support gradient is difficult to achieve and the apparatus and the required force will be

larger than indicated by the equations. Also, it is necessary to develop the necessary gasket support at all chamber pressures higher than S as the pressure is raised and lowered.

Instead of using two moving pistons to compress material inside a relatively stationary cylinder, one can replace the cylinder with an array of pistons with gaskets between them so that the entire apparatus consists of pistons moving toward each other. The simplest example of this device has four pistons arranged tetrahedrally. Six pistons can form a cube or a rectangular parallelepiped. The pistons can be driven by separate rams or driven into tapered bearing rings, or the entire set, suitably sealed, can be immersed in a pressurized liquid. The rewards for this complexity are higher chamber pressures (especially on very compressible materials), a greater number of independent electrical connections (one per piston), and more nearly isotropic compression of the chamber contents.

If we assume that the pistons of a multiple piston apparatus are relatively incompressible compared with the high-pressure-chamber contents or the gaskets, then the chamber volume V is compressed because the gaskets are compressed and/or extruded. If we exclude extrusion for the moment and regard the chamber as a sphere of radius r , its surface area is made of two parts: the area taken by compressing gaskets, $a = 4\pi r^2 f$, where f changes with r , and the area taken by incompressible piston faces, $A = 4\pi r^2 - a$. For a change dr , $dA/dr = 0$, which yields:

$$da/dr = 8\pi r \quad (8)$$

From this, we find:

$$d \ln a = 2dr/rf \quad (9)$$

and the ratio of compressions of chamber and gaskets is

$$d \ln V / d \ln a = 3rf dr / 2r dr = 1.5f \quad (10)$$

For the chamber contents, $d \ln V = K d \ln p$ and for the gasket material, $d \ln a = k d \ln p$, where K and k are compressibilities and p is the same in both the chamber and the gasket next to the chamber (to avoid overstressing the pistons), thus

$$d \ln V / d \ln a = K/k = 1.5f \quad (11)$$

which tells us what the ratio K/k must be to make the pressure in the chamber increase as fast as the pressure in the gasket next to the chamber for any value of r , f , or p .

For a cylinder with fixed ends and shrinking radius, the analog of Eq. (11) is

$$d \ln V / d \ln a = K/k = 2f \quad (12)$$

One might expect f to be about 0.2 and this tells us that K/k should be 0.3 or 0.4 for the sphere or cylinder, respectively. So suitable gasket material should be about

2.5 to 3.3 times as compressible as the chamber contents, assuming that the gasket does not extrude and that the gasket compresses much more than the piston. This is more difficult to achieve at high pressures where differences in compressibilities among various materials become smaller (because the imposed pressure greatly exceeds the normal cohesive internal pressure, so that the interatomic forces are largely repulsive and increase rapidly with decreasing interatomic distance).

At higher pressures (above ~ 8 GPa), the pistons are really not incompressible and their deformation becomes important. This is equivalent to making the f in Eq. (12) somewhat larger and reducing the ratio of k/K required for gasket material. If the pistons are not subjected to excessive stress gradients and deform elastically (reversibly), all is well. However, the chamber pressure then becomes more sensitive to volume changes resulting from phase changes.

Gasket materials are chosen primarily on the basis of their internal friction or extrusion resistance under pressure, compressibilities, thermal stabilities, chemical inertness, and ease of fabrication. Often several materials comprise the gaskets of a particular apparatus. Suitable materials include certain kinds of slightly porous natural stone (e.g., pyrophyllite, or various ceramic materials, and some metals such as copper and steel). Usually slippery or organic materials are unsatisfactory because they tend to fail catastrophically in shear and thereby allow violent extrusion of the gasket and part of the chamber contents.

The pistons of such apparatus tend to slide out of alignment laterally, because this motion increases the chamber volume and reduces the pressure. Conversely, pressure can be generated by forcing offset pistons into alignment.

Under certain conditions, gaskets can extrude. Consider the extrusion of a gasket between two faces diverging at a half angle Θ , as shown in Fig. 4. The gap diverges from $2g$ at $x = 0$. At distance x , the gap has width $2g(1 + x \tan \Theta)$. Let the coefficient of friction of the gasket material be c . If we assume that the ordinary compressive strength of the gasket material is negligible compared with the local pressure p , then the extruding force acting on a slice dx is $-g(1 + x \tan \Theta) dp$, where the minus sign

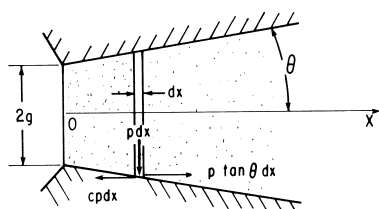


FIGURE 4 Cross section of gap containing softer material being extruded, with local pressure $p(x)$.

indicates that pressure falls as x increases. This force is opposed by wall friction $cp dx$ and assisted by the component due to the slope, $p dx \tan \Theta$. At force balance,

$$-g(1 + x \tan \Theta) dp = (c - \tan \Theta) p dx \quad (13)$$

for which the solution is

$$\ln(p/p_0) = \left(\frac{c - \tan \Theta}{g \tan \Theta} \right) \ln \left(\frac{1 + x_0 \tan \Theta}{1 + x \tan \Theta} \right) \quad (14)$$

where p_0 is the pressure at $x = x_0$. For $\Theta = 0$, Eq. (14) simply becomes:

$$\ln(p/p_0) = (c/g)(x_0 - x) \quad (15)$$

During extrusion the pressure in the gasket falls exponentially with distance, much steeper than ideal (logarithmic) for piston support. One can temper the steepness to some extent by the choice of g or Θ . Note that g falls and the gradient steepens as the chamber pressure rises. This acts to limit extrusion. However, as the pressure is released, extruded material does not flow back toward the high-pressure zone and the pressure gradient in the gasket may become too steep to support the piston properly or prevent extrusion, perhaps violent, of the chamber contents. Thus, extrusion is useful only for pressures below about 6 GPa or during the early stages of compression to higher pressures.

The foregoing comments on the use of compressible gaskets indicate that while in principle a supported piston can withstand a pressure many times its yield strength (the perfect example being a conical column running from Earth's surface to its center), in practice the attainable pressure is limited by problems of size and cost and the relative weakness of piston materials compared with the pressures we would like to reach, as well as by limited knowledge of the compressibilities and rheological properties of candidate gasket materials at higher pressures. Nevertheless, such apparatus can be used for studies of magnesium-iron silicates at pressures of 27 GPa and temperatures of 1000°C .

The sliding anvil scheme avoids the use of gaskets. Figure 5 illustrates the idea. The force F drives the large pistons toward each other and compresses the chamber

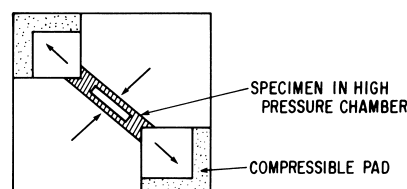


FIGURE 5 Cross section of sliding piston apparatus. The large pistons and the compressible pads are contained by other members not shown.

contents while the smaller pistons slide outward against a confined compressible pad. The compressibility of this pad is matched to the chamber contents to provide piston support. Pressures of over 20 GPa can be produced in this way, but the large changes in the chamber shape are a problem, along with the mechanical complexities and the difficulties of insulating the pistons, particularly at higher temperatures. Hence, the apparatus has not been widely used.

The highest static pressures, over 250 GPa, can be generated in the tiny disk trapped between the faces of two supported anvils made of the strongest piston material—diamond, as shown in Fig. 6. A suitable gasket material is 18–8 stainless steel. The pressurized disk is typically about 0.3 mm in diameter and 0.1 to 0.01 mm thick. The anvil faces must be carefully aligned parallel. A small spring and lever system furnishes the compressive force; the entire apparatus fits easily in a coat pocket. The diamonds are high quality and transparent and permit a direct view or spectral measurements on the compressed material. X-ray diffraction studies can also be made at pressure. Pressure is conveniently measured by the shift of the R_1 ruby line excited by a laser, or by the X-ray diffraction measurement of the lattice spacing of reference substances such as NaCl. The pressure is stable for many days. The entire apparatus can be heated to moderate temperatures or cooled cryogenically. Portions of the compressed material can be briefly heated by laser pulses. Much work significant for geology has been done in this kind of apparatus. The quality of the diamonds and their deformation modes at very high pressures appear to be the principal limitations on

this type of apparatus, which has so markedly extended the understanding of Earth's interior.

Generally, the contents of high-pressure chambers are solid and hence support shearing stresses (i.e., the pressure inside the apparatus is not hydrostatic but varies with position and direction). Usually materials with low shear strength are preferred, such as NaCl, MgO, talc, glasses, and pyrophyllite. The latter three are also moderately good insulating materials for internal furnaces. Hexagonal boron nitride, thoria, and zirconia can be used for extremely high temperatures. Furnace heating elements can be made of refractory metals or graphite. The choices and combinations of materials become limited and specific for the system being studied when temperatures above 2000°C at pressures of 10 GPa or more are used.

The pressure inside the heated chamber may also vary as a result of the local density changes produced by thermal expansion or phase changes resulting from the heating. For example NaCl may expand, melt, and thereby increase the local pressure, while pyrophyllite, a layer-lattice-type aluminum silicate, may transform into a denser assembly of coesite and kyanite, thereby reducing the local pressure. It follows that experimental results in high-pressure, high-temperature work must be interpreted with care.

Thermocouple electromotive force is affected by pressure. Usually the indicated temperature is slightly less than the true temperature. For example at 5.0 GPa and 1300°C, the Pt/Pt10Rh thermocouple indicates a temperature being about 50°C too low. The corrections are approximately linear with both pressure and temperature.

Static pressures above about 3 GPa become increasingly difficult to generate, measure, or estimate because the necessary gasketing support and the pressure gradients inside the chamber complicate the relationship between applied force and generated pressure. Therefore, calibration methods used are based on the behavior of standard materials whose properties change with pressure in known or agreed-upon ways, independently of the type of container. Table III gives the property change and pressure range for some of these reference materials.

For many high-pressure syntheses, the chamber is calibrated cold but used hot and the uncertainties mentioned earlier creep in. Recourse may be had to certain phase transitions that produce characteristic crystalline substances under certain conditions of pressure and temperature. These substances, or evidence for their existence, may then be recovered after cooling and pressure release to indicate the conditions achieved. Of course, the necessary solvents and catalysts must be present to ensure that the transition proceeds as easily as possible at the high temperature.

Some useful transitions of this type and the necessary conditions involved are given in Table IV. The numbers in

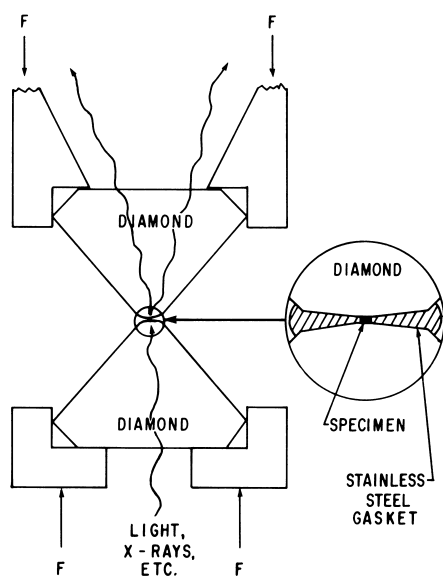


FIGURE 6 Cross section of diamond anvil apparatus with small central face and tapered flanks, for $P \geq 20$ MPa.

TABLE III High Pressure Reference Points

Material	Pressure (GPa)	Property change
Manganin	0.1–10	Electrical resistance; $P = 43.3\Delta R/R_0$ GPa
Bismuth	2.47, 7.75	Phase changes producing changes in density and electrical conductivity
Barium	5.5, 12	Phase changes producing changes in density and electrical conductivity
Lead	13	Phase changes producing changes in density and electrical conductivity
Iron	11	Phase changes producing changes in density and electrical conductivity
Tin	9.6	Phase changes producing changes in density and electrical conductivity
Ruby	1–180	Shift of R_1 fluorescence; $P = 380.8[(\Delta\lambda/\lambda + 1)^5 - 1]$
NaCl	1–100	Compression measured by X-ray diffraction
GaP	~22	Semiconductor to metallic

parentheses are the pressure and temperature coordinates that mark the ends of a straight boundary line between the two phases across which the transition has been observed to occur. The low-pressure phase is listed on the left. The higher the pressure, the greater the uncertainty.

B. Dynamic Pressure Generation

Extremely high pressures can be developed in a piece of matter by accelerating it rapidly. The necessary energy and momentum are provided by rapidly expanding gases, usually from detonating high explosives. The simplest geometrical situation is a plane shock wave developed in the material either from a plane shock wave generated in an adjacent block of high explosive or from the impact of a flyer plate. Typical shock front velocities are of the order of 10 km sec^{-1} , and the pressures range from 10 to 500 GPa.

TABLE IV High-Pressure, High-Temperature Reference Phase Transitions

Phase pair	Transition aid	End points (GPa, °C)
Sillimanite–kyanite	Water	(0.8, 700)–(3.0, 1700)
Quartz–coesite	Water	(2.7, 650)–(4.5, 2100)
NaCl, liquid–solid	—	(2.5, 1250)–(7.0, 1600)
Graphite–diamond	Mn, Ni, Co, Fe	(4.5, 1200)–(10, 3100)
Coesite–stishovite	Water	(8.5, 450)–(10, 1850)
ZnO (NaCl–wurtzite)	Water, shear	(9.5, 200)–(11.5, 600)
αFe – ϵFe	Shear	(10, 500)–(11.3, 25)
ZnSiO_3 : clinopyroxene–ilmenite	Water	(11, 1000)

In a typical specimen a few centimeters thick, the material behind the shock front remains in a compressed and heated state for several microseconds. This is time enough for millions of vibrations of the atoms in the material. Most shock-generated pressures are high enough that the mechanical strength of the material is of minor importance, and the material may be regarded as a fluid. This effect is sometimes used for explosive forming of metal.

For a plane shock wave in a plate that is so wide that edge effects can be neglected (e.g., aspect ratio of 6 or more), simple relationships hold for the central part of the wave.

Figure 7 is a sketch of a block of matter originally at density ρ_0 , specific volume V_0 , internal energy E_0 , and pressure P_0 . Halfway through it is a plane shock wave moving toward the right at velocity s across which the material jumps from rest to a velocity u . Behind this wave the material traveling at velocity u has density ρ , specific volume V , pressure P , and internal energy E .

If we move along with the shock front, we see that the compressed material behind the front does not move as quickly as the original material, and conservation of mass says:

$$\rho_0/\rho = (s - u)/s = V/V_0 \quad (16)$$

For conservation of momentum, the pressure difference across the front accelerates the material according to

$$P - P_0 = \rho_0 s u \quad (17)$$

For conservation of energy, in unit time the work Pu appears as kinetic and internal energy E of the compressed matter that is being formed at a rate of $\rho_0 s$. So we have

$$Pu = \rho_0 s (E - E_0) + \frac{1}{2} \rho_0 s u^2 \quad (18)$$

By combining these three equations we can obtain the Hugoniot equation:

$$2(E - E_0) = (P + P_0)(V_0 - V) \quad (19)$$

When the shock wave reaches the end of the bar, the entire bar is moving to the right at velocity u . The free end of the bar has nothing to push against, and it begins to expand. If friction losses are negligible, this free surface acquires

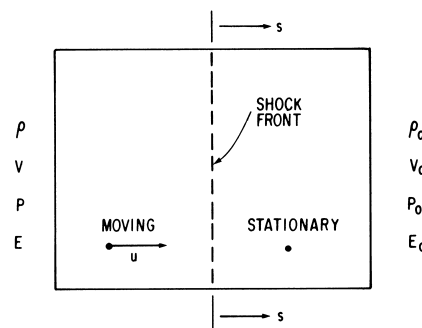


FIGURE 7 Cross section of material being traversed by a plane shock wave of velocity s behind which the material velocity is u relative to the laboratory.

by expansion the same increment of velocity u that it got by compression and it is found that this free-surface velocity is very close to $2u$, relative to the laboratory.

Both $2u$ and s may be measured quite accurately by a variety of techniques such as precisely spaced pins that close electrical circuits and high-speed cameras. Then, from Eqs. (16) to (19) and the initial conditions, one can find P , E , and V for the compressed material behind the shock front and the equation of state $E(P, V)$ of the material near the Hugoniot curve. Various other reasonable assumptions ultimately permit fairly accurate determinations of $E(P, V)$ for pressures and densities further removed from the Hugoniot curve. For each value of P and V , a separate experiment producing particular values of s and u is needed.

If a phase (density) transformation occurs during the shock compression process, the concomitant change in V versus P will be detected. If the transition requires a brief time to be completed, a double wave will form, with the faster wave traveling in the compressed but yet untransformed material. Upon pressure release the transformed material may change back into the low-pressure form, and this process, if slightly delayed, will again produce a separate wave.

The compression in the shock wave is not isotropic but essentially uniaxial and involves considerable shearing. Equation (19) tells us that the temperature rise is larger for more compressible materials. Temperatures of a few thousand degrees Kelvin are easily reached in a gas. If a mixture of materials is shock compressed, the behavior of the system becomes more complex due to differing responses of each material to the compression; geometrical arrangement also becomes important. Shocked material may be recovered in specially designed energy-absorbing catchers that slow the material down without damaging it too much by transferring its momentum to an expendable piece. Multiple or reflected shock waves of more complicated geometry may be used to generate extremely high pressures. High-speed X-ray and electrical techniques can be used to study the state of matter during the few microsecond duration of the shocked state.

Shock-wave phenomena are important in meteorite impacts where high-pressure minerals are often formed. Small diamonds useful for lapping and polishing are made commercially by shocking graphite mixed with iron and copper. The metals cool the diamonds before they can transform back to graphite on pressure release.

IV. GENERAL CONSIDERATIONS OF PHASE CHANGES

High pressures favor denser forms by the term $p\Delta V$ in the system's free energy. Naturally, the denser forms fa-

vor higher numbers of atoms coordinated or bonded about each other. Examples are many. Al_2SiO_5 exists as three polymorphs: sillimanite, andalusite, and kyanite. In the first of these, sillimanite (the low-pressure form), all the aluminum ions are in a four-coordination system with oxygen; andalusite, stable at higher pressures, has half the aluminum ions surrounded by five nearest neighbors; and in kyanite, the highest pressure form, the aluminum ions all have the highest coordination number of 6 with respect to oxygen. In the ordinary forms of silica, quartz, or cristobalite, four oxygen atoms surround each silicon atom; in the high-pressure modification, stishovite, the coordination number of oxygen with respect to silicon is 6 as in the rutile structure.

These examples bring up a second rule: The larger the atom or ions involved, the lower the pressure required for high-coordination numbers. The series carbon-silicon-germanium-tin illustrates this rule very nicely. At low pressures, the coordination number of carbon is 3 (graphite); of silicon, germanium, or gray tin, 4 (diamond structure); and of white tin, 6. At high pressure carbon takes the diamond structure (5 GPa), silicon and germanium take the white tin structure (10 GPa), and white tin changes to a body-centered tetragonal form with coordination number 8. A corollary of the second rule is that the high-pressure forms of lighter elements or compounds are suggested by the low-pressure forms of chemically analogous heavier elements or compounds. This rule is helpful in geological studies.

Among organic materials, reactions that favor higher average density are strongly favored by very high pressures, so that polymerizations, oligomerizations, or ring formations occur easily. The practical difficulty is that most organic molecules, being rather large, stringy, or angular, readily solidify under pressure. In order to mobilize them and achieve reasonable reaction rates, temperatures approaching or exceeding thermal decomposition temperatures may be needed. Thus, very high pressures have been useful mainly for studying small organic molecules or for the elucidation of reaction mechanisms.

As mentioned earlier, atoms forced closer together by very high pressure may also adopt new electronic arrangements. Electrical behavior becomes more metallic as electrons are shared among more atoms. For example, the electrical resistivities of iodine, selenium, sulfur, GaAs, GaP, silicon, germanium, and similar insulators or semiconductors, including large aromatic molecules such as pentacene, fall by many orders of magnitude by the application of 15–50 GPa at 25°C.

In recent years, low-temperature Bridgman anvil apparatus combined with modern instrumentation has led to extensive studies of the superconducting state for pressures up to 50 GPa at temperatures down to 0.05 K. It is found that the 1-atm superconducting substances, such as lead,

mercury, and tantalum, generally show a fall in their superconducting critical temperatures as pressure increases. Interesting exceptions are lanthanum, silver, Mo_6Se_8 , and a few other ternary compounds. However, some substances require pressure for superconductivity—for example, antimony, arsenic, barium, yttrium, germanium, or cesium in their high-pressure forms. The critical temperatures of the latter at first rise with increasing pressure but then usually fall at still higher pressures. Phase changes may complicate this simple picture. So far it appears that the superconductivity in nearly all materials can be explained by the BCS type of electron pairing. Increased understanding will follow as X-ray diffraction studies reveal the crystal structures of the various superconducting substances.

V. PRACTICAL USES OF VERY HIGH PRESSURES

Some modern metalworking processes use very high pressures in extrusion or cold-forming operations simply because the metals being worked are relatively strong and the tools are made of even stronger cemented tungsten carbides. High hydrostatic pressures, 1–2 MPa, have been used to form special pieces that could otherwise be formed only by more expensive methods. Usually the high initial and continuing costs of very high pressure equipment make it a last resort.

A. Diamond Synthesis

Very high pressures probably find their widest use in the commercial synthesis of diamond from graphite. The high value of the products makes the effort economically viable, and several tons of industrial diamonds are synthesized each year in dozens of plants throughout the world.

Figure 8, the carbon phase diagram, forms a basis for discussing the processes involved. Ideal graphite has a density of 2.2 and diamond, 3.52, so 1 ml of graphite becomes 0.63 ml of diamond, a relatively large change. Diamond is favored to form at pressures and temperatures where it is stable, but the carbon atoms must be in the proper environment, particularly at the milder conditions.

At temperatures above about 2500 K, thermal agitation alone is usually sufficient to make the stable phase form in a few seconds or less. Diamond can form from molten carbon (4000 K) in a few milliseconds. The pressures required for diamond stability are then upwards of 10 GPa, which are not economic for static apparatus, and the diamond crystals are very small. However, dynamic pres-

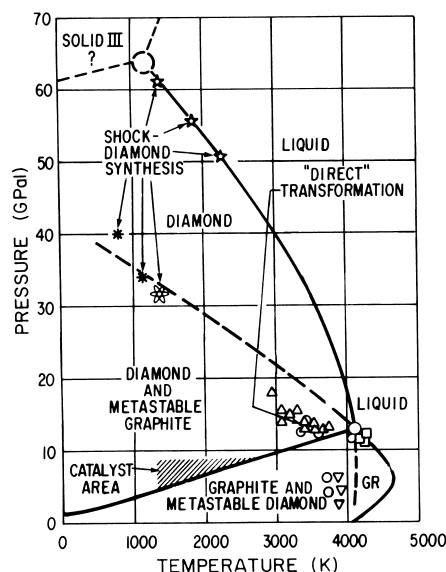


FIGURE 8 Carbon phase diagram showing diamond synthesis regions.

ures of this level are easily reached and tiny diamonds can be made for lapping and polishing, as mentioned earlier.

The bulk of industrial diamond production is done at pressures and temperatures in the range 4.5–6.0 GPa and 1400–1800 K, a range indicated in Fig. 8 by a cross-hatched area. The transformation of graphite to diamond is made possible by using catalyst solvents, which are molten (carbon-saturated) metals such as alloys picked from manganese, iron, cobalt, and nickel. Platinum and palladium are also effective but cost more and require higher temperatures and pressures. (Some carbon solvents, such as AgCl or CdO , do not form diamond from graphite at 5.5 GPa. Diamond-forming catalysts usually carry positively charged carbon in solution.)

Apparatus of the "belt" type is often used. Pieces of graphite and metal occupy the heated zone of the high-pressure chamber. When the chamber pressure has become suitably high, the hot zone temperature is increased until the metal melts and becomes saturated with carbon. At this point, diamond begins to deposit from the molten metal and graphite dissolves. Only a thin layer of metal is involved, and the diamond replaces the graphite. Figure 9 is a photograph of a thin layer of nickel catalyst on the surface of a mass of freshly grown diamonds. The diamonds are recovered after acid treatment.

The higher the pressure over equilibrium, the higher the diamond nucleation and growth rate and the smaller and less perfect the crystal. Lower synthesis temperatures favor cubes and higher ones, octahedra. Suitable control of these variables permits the growth of selected types of



FIGURE 9 Freshly grown diamonds bearing a thin film of nickel. The arrow indicates a bare octahedral face about 0.1 mm in size.

crystals for particular uses (e.g., dendritic, friable crystals cut hard metals more efficiently than blocky crystals, while strong cubo-octahedra are preferred for cutting rock). Commercial diamond grains are now available in sizes up to about 1 mm. Progress in synthesis and application has dramatically reduced the costs of using diamond abrasives since the introduction of synthesized diamond in 1957. The price as of 1986 ranges from about \$5 per gram for powders to about \$15 per gram for 0.7 mm rock-sawing crystals.

The abrasive grains are usually held in the rim of a wheel by a matrix of resin or sintered or electroplated metal. Usually the diamond-bearing part contains 25% or less diamond by volume. The grains themselves may be bare, or they may be precoated with a thin layer of copper or nickel applied by electroless plating techniques (controlled reduction of a solution of metal salt). Diamond surfaces are generally difficult to wet or adhere to, but the metal provides a mechanical grip on the grain while the metal itself is easier to bond to a matrix. As a wheel is used, the diamond grains break up or wear down. In many uses the metal coating helps hold fragments in place and dissipate heat.

The grinding process is more rubbing than cutting, so that local pressures and temperatures are high at the contact areas. Indeed, the processes of abrasion could be considered a branch of high-pressure, high-temperature chemistry, although the conditions are complex and transient. For example, diamond rubbing on clean, hot iron is rapidly attacked, but diamond lasts a long time rubbing or cutting glass. The selection of abrasive grain, type, size, matrix, and so on, depends on the particular application and is usually done by extensive testing under various

conditions. Selection is not done by theory because of the complexities of the grinding process.

The need for industrial diamond in sizes over 1 mm has largely been met by a sintered diamond material made of smaller grains. The sintering process uses pressures and temperatures similar to those for initial synthesis with a few percent by volume of a sintering aid, usually cobalt. The cobalt helps to form direct diamond-to-diamond bonds, which give the mass high hardness and thermal conductivity. The randomly oriented polycrystalline structure gives good strength and shock resistance. Such pieces of sintered diamond are widely used for cutting tools for hard materials, including rock (but not iron, nickel, or cobalt-based alloys), for dies for drawing wire, and for dressing abrasive wheels. The residual cobalt weakens them at temperatures above about 800°C, but those from which the cobalt has been leached, though not as strong, are durable to 1200°C. Various shapes are available in sizes up to about 2 cm. Some have been used for special very high pressure apparatus reaching 50 GPa.

Single diamond crystals of gem quality can be grown at high pressure using the reaction cell arrangement shown in Fig. 10. The carbon source is a mass of small diamond crystals maintained at 1450°C at the top of a bath of molten catalyst metal alloy (e.g., iron). Diamond grows on a seed crystal held at about 1425°C at the bottom of the bath. Stray diamond nuclei tend to float up out of the growing zone. The bath is held in a sodium chloride vessel whose melting temperature is above 1450°C at the operating pressure, about 5.5 GPa. About a week is needed to grow an acceptable single crystal about 1.3 carat and about 6 mm in maximum dimension; recently a 3.5 carat crystal was grown in about 200 hours. The process is not economically practical, but special crystals of scientific interest are grown. Many of these are more pure and more internally perfect than any natural diamond. A few parts per million of nitrogen yield yellow crystals in which nitrogen atoms replace carbon atoms. A few parts per million of boron yield blue, *p*-type semiconducting crystals. Figure 11 shows several crystals of various types.

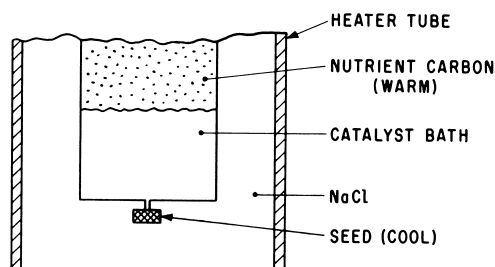


FIGURE 10 High-pressure cell for growing single diamond crystals.

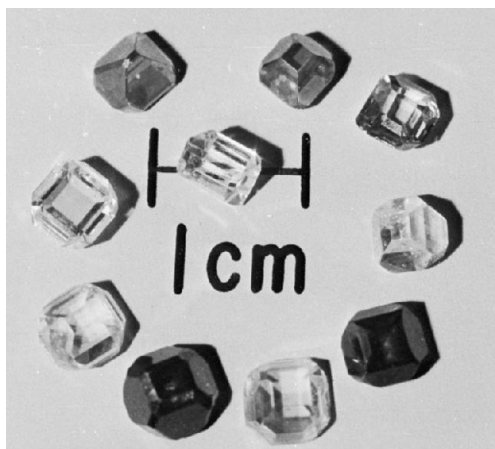


FIGURE 11 Synthesized high-quality diamond crystals, showing their typical growth faces.

Special single diamond crystals containing about 99.9% of the carbon-12 isotope have been grown to about 5 mm in size using the method described above. The carbon-12 source diamond crystals were made by the low-pressure (10 torr) decomposition of carbon-12 methane at about 950°C in the presence of hydrogen atoms generated nearby by a hot tungsten wire. (The H atoms keep the solid carbon surface atoms in tetrahedral bonding states.) These diamonds are noteworthy for their excellent thermal conductivity at 20°C, about 8 times that of copper and 5 times that of most diamonds.

When graphite of good crystalline perfection is compressed to 10–14 GPa and heated to about 1000°C, it mostly collapses into diamond. Much of this diamond is not cubic but hexagonal, like the wurtzite structure. This happens because melting did not occur, and the diamond form was forced by the form of the graphite. The graphite collapsed in a direction parallel with the hexagonal sheets of atoms, like squeezing a deck of cards on the edges, not the faces. Traces of hexagonal diamond, called lonsdaleite, also appear in shock-formed diamond, natural or synthetic. Lonsdaleite is slightly less stable than regular cubic diamond and changes to cubic if heated hot enough or exposed to solvent catalysts at high pressures.

The tendency for diamond formed under nonfluid conditions to be influenced by the structure of the precursor carbon can be noted when hydrocarbons are decomposed at 12 GPa. Aliphatic hydrocarbons, which already possess tetrahedral carbon bonding, seem to slowly lose hydrogen and approach cubic diamond. Purely aromatic molecules such as anthracene change to graphite, then finally to diamond at higher temperatures. Adding aliphatic carbon atoms to the molecules or the mixture favors diamond formation at lower temperatures.

B. Cubic Boron Nitride

Boron nitride, BN, exists in three forms: (1) a hexagonal form, such as graphite; (2) a dense cubic form (zincblende structure, such as diamond); and (3) a dense hexagonal form (wurtzite, such as lonsdaleite). The two dense forms are thermodynamically stable only at higher pressures, but, like diamond, can be formed at high pressures and high temperatures and then quenched and recovered for use or study at atmospheric pressure. The equilibrium pressures for cubic BN are about 10% lower than those for diamond. The dense hexagonal form is slightly less stable than the cubic form and is usually prepared by shock compression or catalyst-free pressure and heat (10 GPa, 1000°C) from crystalline graphitic BN.

Cubic BN is usually manufactured at about 5 GPa and 1500°C from a mixture of graphitic hexagonal BN and a catalyst solvent such as lithium or magnesium nitride. Many other catalyst solvent systems have been found and most of them involve a nitride-forming element. As pressure and temperature increase, the catalyst requirements relax as with carbon.

The cubic form is widely used as an abrasive or as sintered cutting tools for grinding or shaping hard ferrous-, nickel-, or cobalt-based alloys. It is not quite as hard as diamond, but it is more resistant to oxidation and alloying with the workpiece metal. Its low wear rate and cool cutting action make it a favorite for high-precision work on cutting tools, cylinders, and rotors. Its price is similar to that of synthesized diamond. So far, high-quality single crystals up to about 4 mm in size have been grown at high pressures using the temperature-difference technique used with diamond. The bath was an alkaline earth nitride-BN complex contained in a molybdenum can. It was possible to grow *n*-type (S-doped) BN on a *p*-type (Be-doped) BN seed crystal to form a *p-n* junction diode a few millimeters in size which emitted blue light when carrying current in the forward direction.

C. Synthesis of Other Inorganic Materials

Although at least hundreds of new high-pressure phases have been made in the search for other materials with useful applications, the primary benefit has been greater understanding in solid-state chemistry and physics. The closely related effort to understand the properties of the deeper materials of Earth and the other planets will continue to be one of the driving forces for high-pressure studies. Metallic ammonia and metallic hydrogen are of direct interest in the structure of the larger planets, and it is hoped that the conditions for synthesis of metallic hydrogen might be attained in the diamond anvil. It is estimated that above about 300 GPa would be required. This

pressure is still somewhat above the maximum reached in the diamond anvil to date (1985), about 270 GPa. Metallic hydrogen has also been suggested as a room-temperature superconductor, and undoubtedly would be a tremendous propellant if it could be brought back alive, although this is considered to be unlikely. The best evidence to date for the existence of metallic hydrogen is from shock experiments.

Another candidate for a useful material from very high pressure synthesis is the gem material, jadeite ($\text{NaAlSi}_2\text{O}_6$). The natural material of "Imperial" quality can cost as much as \$2000 per carat. Jadeite can be synthesized at about 30 kb and above in equipment similar to that used for diamond growth, and it has been made into pieces of jewelry. Since jadeite is used as a polycrystalline aggregate, synthesis is essentially hot pressing and sintering, much simpler than if single crystals were needed. However, it does not appear to be a commercial product in competition with the natural supply.

If one considers the "low" pressure range around 0.1–0.2 GPa, there are two economically viable syntheses: that of the quartz form of SiO_2 and the magnetic tape material, CrO_2 . The latter is made from the decomposition of CrO_3 in a steel vessel where the oxygen pressure is maintained at about 0.03 GPa and the temperature is around 400°C . This pressure is necessary in order to maintain the Cr^{4+} state. The acicular grains that form are used for high-resolution tape recording.

The quartz form of SiO_2 is a very important piezoelectric material used for transducers and frequency control devices. This industry used to be dependent on small, often unreliable sources from sometimes unstable political environments (e.g., Brazil) for optical-grade natural quartz, and the useful yield from this material was variable. Now many large crystals are grown in plants all over the world on seeds in large pressure vessels from the system $\text{Na}_2\text{O}-\text{SiO}_2-\text{H}_2\text{O}$ with a temperature gradient under much the same conditions nature is thought to have used: about 1 kbar and 500°C . This combination produces a well-controlled useful product compatible with production methods and essentially frees the industry from dependence on natural material. Natural mica, topaz, asbestos, and other OH-containing minerals also grew under similar hydrothermal conditions in nature's pressure vessels, such as pegmatite dikes. This part of the universe, including the genesis of ore bodies in Earth's crust, is a continuing area of moderately high-pressure investigations. There has also been considerable success in growth of a variety of crystals other than quartz using the hydrothermal method, but in general a simpler alternative method is sought even in the 0.1- to 0.2-GPa range (e.g., emerald can be grown hydrothermally, but solution growth at high temperatures at 1 atm is preferred).

VI. JOURNEY TO THE CENTER OF THE EARTH

On the basis of seismological data from earthquakes, Earth's interior, in broad terms, consists of a crust, a mantle, and a core. The crust and mantle are primarily oxides and the core is primarily metallic. The details of this structure are continually being updated and redefined as techniques become more sophisticated. Although Earth's interior is essentially inaccessible by direct observations (the deepest well is only 12 km), there has been considerable help in modeling the interior from high-pressure research, principally by studying properties and phase transformations of the known and surmised oxide minerals. Standard static apparatus plus the diamond anvil have been used to establish the P and T conditions for stability of possible phases and to measure their densities and establish crystal structures and seismic properties under pressure. Figure 12 shows the extent to which the interior of Earth has been probed in the laboratory by high-pressure and high-temperature studies.

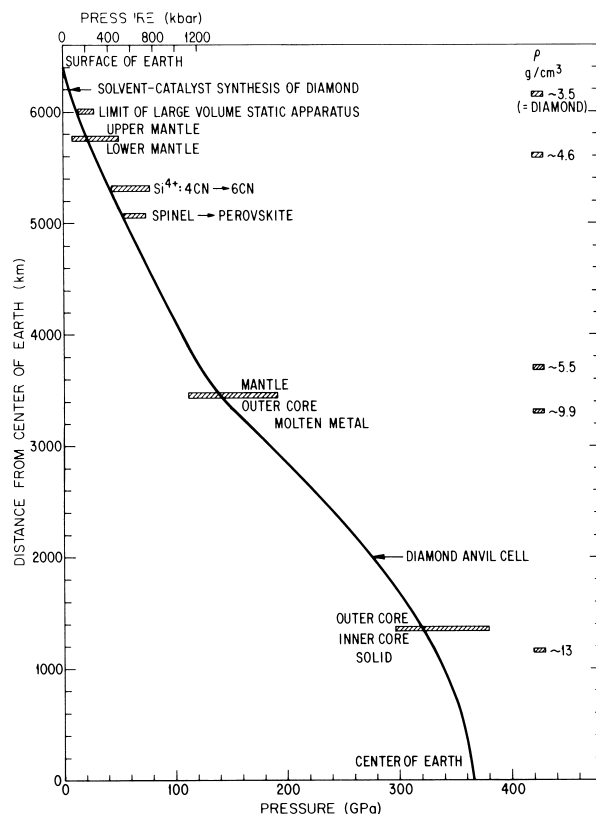


FIGURE 12 Pressure versus depth in Earth. Crust, mantle, and core boundaries and densities are indicated along with pressures attainable with the diamond anvil and large-volume static apparatus.

Up to where these materials dissociate, the obvious effect of high pressures (depth in the earth) is increased density, which is accomplished structurally by atomic rearrangements in the crystal lattice. The principal coordination changes for aluminum and silicon (both with four to six nearest oxygen neighbors) have been mentioned for important minerals such as aluminum silicates and quartz. Other very important phases are represented by pyroxene (MgSiO_3) and forsterite (Mg_2SiO_4), both of which are common in the basic igneous rocks of the upper mantle and crust. Changes in forsterite include transformation to a spinel phase of the same composition and then disproportionation to MgSiO_3 and MgO at about 700 km. The MgSiO_3 phase transforms to an ilmenite structure and then to a perovskite lattice without composition change. This means a change in the coordination number of silicon from 4 in the 1-atm pyroxene form to 6 in the other forms. The magnesium coordination number also increases as these structural changes take place. Seismic velocity changes would be expected at the zone boundaries representing these transitions, but the demarcation may be fuzzy because of composition gradients and substitution of other ions in these structures (e.g., Fe for Mg).

Forsterite (Mg_2SiO_4) is a constituent of a most interesting and mysterious rock, kimberlite, which is the host of natural terrestrial diamond, although only a small percentage of kimberlites contain diamond and fewer yet in amounts warranting mining. It is still controversial whether diamonds are formed in the kimberlite or are simply carried into their present locations by this igneous rock. In any case, diamonds in kimberlite often contain inclusions of the following minerals: forsterite (a form of olivine), pyroxene, garnet, the coesite form of SiO_2 (without the stishovite form), and others. This obviously means these phases are present as small crystals simultaneously with the growing diamond. By determining the pressure and temperature conditions for their stability, it is possible to bracket the conditions for diamond synthesis in Earth's mantle. Thus, from laboratory studies, diamonds are apparently formed at depths of about 100 to 300 km (about 3.5 to 10 GPa) and temperatures above 1000°C . The upper pressure limit is based on the fact that coesite but not stishovite is found in kimberlites. These limits are surprisingly close to those found in the metal-carbon systems from which diamond is manufactured (e.g., 4 to 6 GPa). This agreement is a bit surprising because, while metal inclusions are common in manufactured diamonds, there is no evidence of elemental metal as an inclusion inside natural diamonds from kimberlites, so the chemistries of the two growth systems differ.

Several polycrystalline varieties of diamond exist, ranging from somewhat porous or contaminated masses, such as framesite or carbonado, to ballas, which is essentially pure carbon. Ballas is found only in northwestern Africa,

Brazil, and Russia. The simplest diamonds to understand are the small, dark fine-grained fragments, found in a few meteorites, which undoubtedly formed from graphite by shock compression and heating during impact.

Most natural diamonds are dark or flawed. Especially puzzling are red and brown hues. Even the colorless crystals, when sectioned and examined by fluorescence, etching, and other techniques, reveal many layers of growth. Isotopic dating methods indicate that most diamonds are several thousand million years old.

Another characteristic of natural diamond is its nitrogen content. Most, called type Ia, have many parts per million of nitrogen in the form of coalesced groups of nitrogen atoms. They produce an infrared absorption at 1280 cm^{-1} but are inactive in electron paramagnetic resonance (EPR). The more rare type Ib diamonds are yellow due to isolated nitrogen atoms that replace carbon atoms. They absorb light at 1130 and 1343 cm^{-1} in the infrared and show an EPR spectrum. After an hour in the laboratory at 1800 – 1900°C and 6 GPa, the type Ib were largely transformed to type Ia; most of the nitrogen atoms had coalesced. Synthesized type Ib diamonds behaved similarly. Evidently natural type Ib diamonds did not experience temperatures above about 1500°C for more than a year.

Most of the kinds of natural diamond have not yet been duplicated in the laboratory. In fact, an active area of research is the high-pressure, high-temperature chemistry of carbon in rocks at depth in the earth and how it got together to form diamond crystals. The studies seem to center on the system C–H–Si–O with the possibility of species such as CH_4 , CO , and CO_2 .

With increasing depth in Earth, the oxide compounds tend to dissociate to simpler oxides and finally only metal alloys are stable at the core. The metal-rock boundary at the core is quite distinct. The density of Earth's metallic core at the pressures known to exist there indicate that it contains a significant fraction of elements lighter than iron. If diamond anvils can be improved, some incremental progress in the observation and interpretation of these trends may be expected.

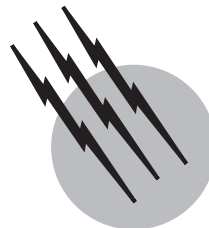
SEE ALSO THE FOLLOWING ARTICLES

CHEMICAL THERMODYNAMICS • EARTH'S CORE • HIGH-PRESSURE RESEARCH • SUPERCONDUCTIVITY

BIBLIOGRAPHY

- Ahrens, T. J. (1980). "Dynamic compression of earth materials," *Science* **207**, 1035–1040.
- Anthony, T. R. *et al.* (1990). "Carbon-12 enriched diamond with high thermal conductivity," *Phys. Rev. B* **42**, 1104.

- Bell, P. M., Mao, H. K., and Goettel, K. A. (1984). "Ultrahigh pressure: beyond 2 megabars and the ruby fluorescence scale," *Science* **226**, 542–544.
- Bridgman, P. W. (1949). "The Physics of High Pressure," Bell, London (Dover, New York, 1970).
- Bundy, F. P., Strong, H. M., and Wentorf, R. H., Jr. (1973). "Methods and mechanisms of synthetic diamond growth," In "Chemistry and Physics of Carbon" (P. L. Walker, Jr. and P. A. Thrower, eds.), Vol. 10, pp. 213–263, Marcel Dekker, New York.
- Homan, C., MacCrone, R. K., and Whalley, E., eds. (1984). "High pressure in science and technology," Parts I, II, III (*Proc. AIRAPT Int. High Pressure Conf., 9th, Albany*), North-Holland, New York.
- Jayaraman, A. (1984). "The diamond-anvil high pressure cell," *Sci. Am.* **250** (4), 54–62.
- McWhan, D. B. (1972). "The pressure variable in materials research," *Science* **176**, 751–758.
- Mishima, O. *et al.* (1987). "Cubic BN light-emitting diode," *Science* **238**, 181–183.
- Pistorius, C. W. F. T. (1976). "Phase relations and structures of solids at high pressures," *Prog. Solid State Chem.* **11**, 1–151.
- Rigden, S. M., Ahrens, T. J., and Stolper, E. M. (1984). "Densities of liquid silicates at high pressure," *Science* **226**, 1071–1074.
- Sunagawa, I., ed. (1984). "Materials Science of the Earth's Interior," Terra Scientific Publ., Tokyo; Reidel Publ., Dordrecht, Holland.
- Wentorf, R. H., Jr., ed. (1974). "Advances in High-Pressure Research," Vol. 4, Academic Press, New York.
- Wentorf, R. H., Jr., DeVries, R. C., and Bundy, F. P. (1980). "Sintered superhard materials," *Science* **208**, 873–880.



Mass Transfer and Diffusion

E. L. Cussler

University of Minnesota

- I. Diffusion
- II. Dispersion
- III. Mass Transfer
- IV. Conclusions

GLOSSARY

Convection bulk flow, usually the result of forces on the system, but occasionally caused by diffusion.

Diffusion spontaneous differential mixing caused by Brownian motion.

Diffusion coefficient the flux divided by the concentration gradient.

Dispersion spontaneous mixing effected by flow and—only sometimes—by diffusion.

Flux the moles or mass transported per area per time.

Mass transfer spontaneous mixing from a system's boundary into its bulk.

Mass transfer coefficient the flux divided by the concentration difference between an interface and the bulk.

IF A FEW CRYSTALS of blue copper sulfate are placed in the bottom of a tall bottle filled with water, the color will slowly spread through the bottle. At first, the color will be concentrated in the bottom. After a day, it will penetrate upward a centimeter or so. After several years the solution will appear to be homogeneous.

The process responsible for the movement of the copper sulfate is diffusion, the basic phenomenon in this article. Caused by random molecular motion, diffusion leads

to complete mixing. It is often a slow process. In many cases diffusion occurs sequentially with other phenomena. When it is the slowest step in the sequence, it limits the overall rate of the process.

In gases and liquids, the rates of these diffusion processes can often be accelerated by convective flow. For example, the copper sulfate in the tall bottle can be completely mixed in a few minutes if the solution is stirred. This accelerated mixing is not due to diffusion alone, but to a combination of diffusion and convection. Diffusion still depends on the random molecular motions that take place over small molecular distances. The convective stirring is not a molecular process, but a macroscopic process which moves portions of the fluid over longer distances. After this macroscopic motion, diffusion mixes the newly adjacent portions of the fluids.

The description of diffusion involves three complementary mathematical models, often dignified as “laws.” The most fundamental, Fick’s law of diffusion, uses a “diffusion coefficient.” In other cases, where convection is strong, the mixing will occur following the same mathematics as Fick’s law but with a “dispersion” coefficient replacing the diffusion coefficient. In still others cases, where there is transport across some type of interface, the mixing is described as “mass transfer” and correlated with a “mass transfer coefficient.” Mass transfer coefficients

provide the basic description of commercial separation processes and hence supply an important topic of chemical engineering.

Choosing between these three approaches is not always easy. Diffusion problems normally give a concentration profile as a function of position and time. Dispersion can do the same, but dispersion tends to be dependent solely on the physics, and not be affected by chemistry. Mass transfer coefficients, on the other hand, tend to describe concentrations as a function of position or time, rather than both variables at once.

In general, diffusion is most useful for fundamental studies where we want to know the details about the system. For example, if we were concerned with a plasticizer inside a polymer film, we might want to know where and when the plasticizer is located. Diffusion will tell us. Dispersion can be important when there is convection, as in chromatography or atmospheric pollution. Mass transfer, on the other hand, tends to be useful in less fundamental, more practical problems. For example, if we want to know how to humidify and ventilate a house, we probably will use mass transfer coefficients.

We will emphasize diffusion and mass transfer in this article, for these are two of the more important processes in chemical engineering. We will mention dispersion simply because insights into diffusion are often a valuable aid in understanding dispersion. We turn first to the subject of diffusion itself.

I. DIFFUSION

A. Basic Equations

The key equation describing diffusion, commonly called Fick's law, asserts that the flux, that is, the amount of solute per area per time, is proportional to the concentration gradient, that is, the derivative of the concentration with respect to position (Graham, 1850 and Fick, 1855). In quantitative terms, this relationship in one dimension can be written as

$$-j_1 = D \frac{dc_1}{dz} \quad (1)$$

where j_1 is the flux in, for example, moles per area per time; c_1 is the concentration in, for example, moles per volume; z is the position, and D is a proportionality constant called a diffusion coefficient. In three dimensions, this can be written as

$$-j_1 = D \nabla c_1 \quad (2)$$

which recognizes that the flux is a vector and the concentration can vary in all three dimensions. In this article we will almost always restrict our discussion to one-dimensional diffusion because this is the most important case and the easiest to understand. Problems in-

volving diffusion in many dimensions are treated in detail elsewhere (Crank, 1975 and Carslaw *et al.*, 1986).

B. Diffusion Across a Thin Film

We can explore the use of Fick's law by considering three key cases (Cussler, 1997). The easiest case for this variation occurs across a thin film like that in Fig. 1. In this figure, we show one large well-stirred volume of a fluid containing a solute at concentration, c_{10} . It is separated by a thin film from another well-stirred volume of solution at a different concentration, c_{1l} . We want to find how this concentration varies between these two volumes.

To find this variation, we make a mass balance on a thin layer Δz thick located at some arbitrary position z within the thin film. The mass balance on this layer is

$$\text{solute accumulation} = \text{diffusion in-out} \quad (3)$$

Because the volumes adjacent to the film are large, the process is in steady state and the accumulation is zero. The mass balance is thus

$$0 = j_1|_z - j_1|_{z+\Delta z} \quad (4)$$

Dividing by Δz and taking the limit as Δz goes to zero, we obtain

$$0 = -\frac{dj_1}{dz} \quad (5)$$

When we combine this with Fick's Law, we get

$$0 = D \frac{d^2 c_1}{dz^2} \quad (6)$$

This is subject to the boundary conditions

$$z = 0 \quad c_1 = c_{10} \quad (7)$$

$$z = l \quad c_1 = c_{1l} \quad (8)$$

The result is easily integrated to find the concentration profile:

$$c_1 = c_{10} - (c_{10} - c_{1l})z/l \quad (9)$$

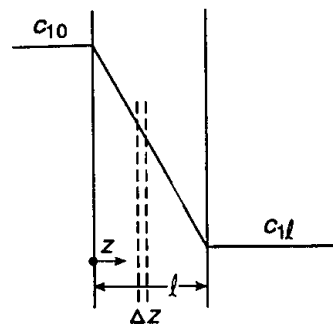


FIGURE 1 Diffusion across a thin film. This is the simplest diffusion problem, basic to perhaps 80% of what follows. Note that the concentration profile is independent of the diffusion coefficient.

This concentration profile can now be put back into Fick's Law to find the flux across the thin film:

$$j_1 = \frac{D}{l}(c_{10} - c_{1l}) \quad (10)$$

This result says that the concentration profile is linear, as implied by Fig. 1. It says that the flux will double if the diffusion coefficient is doubled, if the concentration difference across the film is doubled, or if the thickness of the film is cut in half. This important result is often undervalued because of its mathematical simplicity. However, anyone wishing to understand this subject should make sure that each step of this argument is understood.

C. Diffusion into a Semi-Infinite Slab

The second key case for diffusion occurs when the diffusion takes place not across the thin film but into a huge slab which has one boundary at $z = 0$. In this case, shown schematically in Fig. 2, the concentration is suddenly raised at time zero from $c_{1\infty}$ to c_{10} . As a result, the concentration changes as shown in the figure. We want to calculate this concentration profile.

As before, we start with mass balance written on a thin layer Δz thick:

$$\text{solute accumulation} = \text{diffusion in-out} \quad (11)$$

This situation is an unsteady state, so there is solute accumulation. By arguments that parallel those which let us go from Eq. (4) to Eq. (6), we now get the result

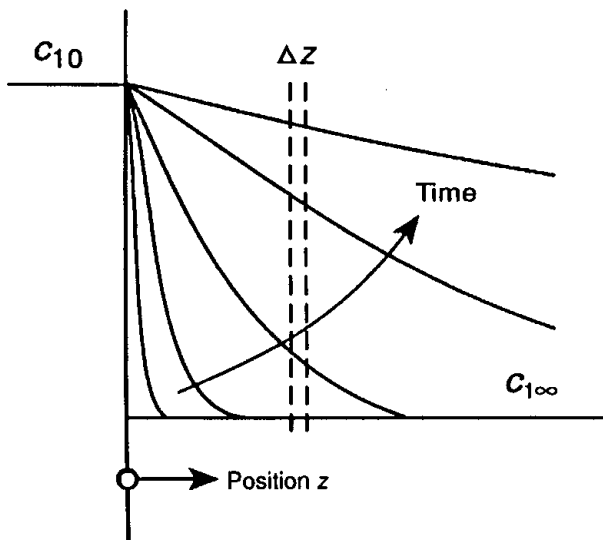


FIGURE 2 Free diffusion. In this case, the concentration at the left is suddenly increased to a higher constant value. Diffusion occurs in the region to the right. This case and that in Fig. 1 are basic to most diffusion problems.

$$\frac{\partial c_1}{\partial t} = D \frac{\partial^2 c_1}{\partial z^2} \quad (12)$$

This is subject to the constraints

$$t = 0 \quad \text{all } z \quad c_1 = c_{1\infty} \quad (13)$$

$$t > 0 \quad z = 0 \quad c_1 = c_{10} \quad (14)$$

$$z = \infty \quad c_1 = c_{1\infty} \quad (15)$$

This case of the semi-infinite slab can be solved to yield both a concentration profile and an interfacial flux which are

$$\frac{c_1 - c_{10}}{c_{1\infty} - c_{10}} = \text{erf} \frac{z}{\sqrt{4Dt}} \quad (16)$$

$$j_1|_{z=0} = \sqrt{\frac{D}{\pi t}}(c_{10} - c_{1\infty}) \quad (17)$$

where $\text{erf}(x)$ is the error function of x . These two equations represent the second key case of diffusion. While they are probably ten times less important than Eqs. (9)–(10), they are more important than any other solutions of diffusion problems.

D. Diffusion of a Pulse

The third key case for diffusion occurs when the solute is originally present as a very sharp pulse, like that shown in Fig. 3. The total amount of material in the pulse is M and the area across which the pulse is spreading perpendicular to the direction of diffusion is A . Under these cases the concentration profile is Gaussian:

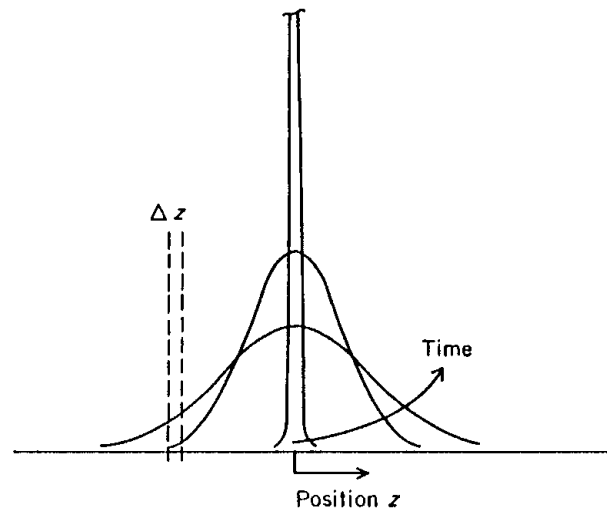


FIGURE 3 Diffusion of a pulse. The concentrated solute originally located at $z = 0$ diffuses as the Gaussian profile shown. This is the third of the three most important cases, along with those in Figs. 1 and 2.

TABLE I A Comparison of Diffusion Coefficients and Their Variations

Phase	Typical value			Variations		Remarks
	cm ² /sec	Temperature	Pressure	Solute size	Viscosity	
Gases	10 ⁻¹	$T^{3/2}$	p^{-1}	(Radius) ⁻²	μ^{+1}	Successful theoretical predications
Liquids	10 ⁻⁵	T	Small	(Radius) ⁻¹	μ^{-1}	Can be concentration-dependent
Solids	10 ⁻¹⁰	Large	Small	(Lattice spacing) ⁺²	Not applicable	Wide range of values
Polymers	10 ⁻⁸	Large	Small	(Molecular weight) ^(-0.5 to -2)	Often small	Involve different special cases

Note: These heuristics are guides for estimates, but will not always be accurate.

$$c_1 = \frac{M/A}{\sqrt{4\pi Dt}} e^{-\frac{z^2}{4Dt}} \quad (18)$$

In fact, this particular problem is not that important for diffusion itself but as the basis of dispersion, discussed below. As a result, we defer further discussion for now.

E. Diffusion Coefficients

So far, we have treated the diffusion coefficients which appeared above as parameters which would necessarily need to be determined by experiment. As a result of 150 years of effort, the experimental measurements of these coefficients are now extensive. Their general characteristics are shown in Table I (Cussler, 1997). In general, diffusion coefficients in gases and liquids can be accurately estimated, but those in solids and polymers can not. In gases, estimates based on kinetic theory are accurate to around 8%. In liquids, estimates based on the assumption that each solute is a sphere moving in a solvent continuum are accurate to around 20%, but can be supplemented by extensive data and empiricisms (Reid *et al.*, 1997).

Other characteristics are harder to generalize. The typical values given in Table I are reasonable, for the coefficients do tend to group around the estimates given. This is less true for solids than for the other phases. The variation of diffusion coefficients with temperature is large in solids and polymers, but small in gases and liquids. Variations of the coefficients with pressure are small except for gases. Interestingly, the diffusion coefficient is proportional to the viscosity in gases, but is inversely proportional to the viscosity in liquids. Beyond these generalizations, we recommend using data whenever possible.

F. Problems with this Simple Picture

The simple picture of diffusion given above ignores several issues that can be important. These include diffusion-engendered convection, multicomponent diffusion, and the limits of Fick's law. Each of these merits discussion.

We begin with the diffusion-engendered convection. In general, the total flux is the sum of the diffusive flux and the convective flux. For example, imagine we had a cup

of coffee in which we dropped a lump of sugar. We would describe diffusion as how fast the sugar moved within the coffee cup, independent of whether the coffee was on the kitchen table or in an airplane flying at 1000 km/hr. Thus when we are considering diffusion, we would sensibly subtract any additional motion of the system.

But with diffusion, things are not always quite so simple. As an example, consider the basic apparatus shown in Fig. 4 (Cussler, 1997). In this apparatus two identical bulbs contain different gases. For example, the bulb on the left might contain nitrogen and the bulb on the right might contain hydrogen. Because nitrogen's molecular weight is higher, the initial center of mass would be closer to the nitrogen bulb, as shown in the figure. If we now open the valve between the two bulbs and allow diffusion to take place, we will wind up with the two bulbs finally containing equal amounts of hydrogen and of nitrogen. That means that the final center of mass will be in the center

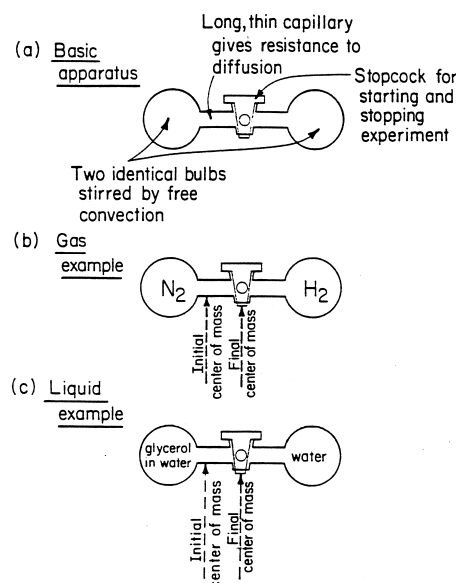


FIGURE 4 An example of reference velocities. Descriptions of diffusion imply reference to a velocity relative to the system's mass or volume. While the mass often has a nonzero velocity, the volume often shows no velocity. Hence, diffusion is best referred to the volume's average velocity.

of the apparatus. Because the center of mass has moved, there must be some convection. Yet we would expect this process to be completely described by diffusion.

In fact, we are right in our expectation. The total flux, the sum of the diffusive flux and the convective flux, can be written as

$$n_1 = c_1 v_1 \quad (19)$$

where c_1 and v_1 are the concentration and velocity of the solute of interest. We can then split off a convective velocity v^0 as:

$$n_1 = c_1(v_1 - v^0) + c_1 v^0 \quad (20)$$

The first term on the right-hand side of this equation is that due to diffusion, so that we can write Eq. (20) as

$$n_1 = j_1 + c_1 v^0 \quad (21)$$

While this much is straightforward, the choice of the velocity v^0 can be complicated, beyond the scope of this article (Taylor *et al.*, 1993 and de Groot *et al.*, 1962). Fortunately, this is not normally significant. When we want a very accurate description, we should consider this additional factor.

In addition to convection, we must recognize that Fick's law applies exactly to only one solute and one solvent, i.e., to a binary system. In general we should write a more complete flux equation like (de Groot *et al.*, 1962 and Katchalsky *et al.*, 1967):

$$-j_i = \sum_{j=1}^{n-1} D_{ij} \nabla c_j \quad (22)$$

which is often referred to as a generalized Fick's law form of multicomponent diffusion equation. For an n component system, Eq. (22) has $(n-1)^2$ diffusion coefficients of which $n(n-1)/2$ are independent. Alternatively, one can use a different form of diffusion equation which for ideal gases is (Taylor *et al.*, 1993):

$$\nabla y_i = \sum_{j=1}^n \frac{y_i y_j}{D_{ij}} (v_j - v_i) \quad (23)$$

where y_i is the mole fraction of species i and the D_{ij} here are the binary diffusion coefficients. This equation, frequently called the Maxwell-Stefan form, is attractive intellectually but can be hard to use. Fortunately, the entire subject of multicomponent diffusion is not that important because any solute present at high dilution will follow the binary form of Fick's law.

The final issue is the validity of Fick's law itself. On the basis of irreversible thermodynamics (Taylor *et al.*, 1993; de Groot *et al.*, 1962; and Katchalsky *et al.*, 1967), one can show that an alternative form of Fick's law is

$$-j_1 = \frac{D_0 c_1}{k_B T} \nabla \mu_1 \quad (24)$$

where $\nabla \mu_1$ is the gradient of the solute's chemical potential. One ordinarily expects the concentration to vary with chemical potential as

$$\mu_1 = \mu_1^0 + k_B T \ln c_1 \gamma_1 \quad (25)$$

where γ_1 is an activity coefficient. Combining these relationships, we find

$$-j_1 = \left[D_0 \left(1 + \frac{\partial \ln \gamma_1}{\partial \ln c_1} \right) \right] \nabla c_1 \quad (26)$$

This says that the diffusion coefficient should vary with the activity coefficient.

The interesting feature of Eq. (26) is that it predicts the diffusion coefficient will go to zero at a critical point or a consolute point. This is verified experimentally: the diffusion coefficient does drop from a perfectly normal value by more than a million times over perhaps just a few degrees centigrade (Kim *et al.*, 1997). Curiously, the drop occurs more rapidly than predicted by Eq. (26). In many ways, this is a boon, because the diffusion coefficient is small only in a very small region of little practical significance. However, it is disquieting that we do not understand completely why the drop is faster than it should be.

II. DISPERSION

At this point we can benefit from a tangent by discussing dispersion, a different effect than diffusion but described by the same mathematics. Unfortunately, dispersion is frequently called "diffusion" in some literature. As a result, it seems sensible to cover it here, if only to show why the processes are different.

A good example of dispersion is a plume of smoke being swept away by the wind. This plume will normally assume a Gaussian profile, a bell-shaped curve whose width is a function of the dispersion coefficient. If the amount of smoke emitted per time S is a constant, then the concentration of material in the smoke is given by (Seinfeld, 1985)

$$c_1 = \frac{S}{4\pi x E} e^{-\frac{z^2}{4Et}} \quad (27)$$

where x is the distance down wind, E is the dispersion coefficient, z is the direction perpendicular to the wind, and t is the time. This has a similar Gaussian dependence as that found for diffusion of a pulse, shown in Eq. (18).

The dispersion coefficient E shown in Eq. (27) is not equal to the diffusion coefficient defined in the earlier parts of this entry. The dispersion coefficient does have the same dimensions of length² per time as the diffusion coefficient. Its function is to describe how fast the smoke spreads, just as the diffusion coefficient describes how fast the solute spreads. However, the dispersion coefficient E is much more a function of physics and much less a function of chemistry. For example, we expect the diffusion

coefficient of hydrogen sulfide to be different than the diffusion coefficient of hydrogen, because these are two different chemical species. However, the dispersion coefficient of hydrogen sulfide in the smoke will be the same as the dispersion coefficient of the hydrogen in the smoke because the mechanism is not that of molecular motion, but rather of velocity fluctuations.

Dispersion coefficients are usually much greater than diffusion coefficients and cause much more rapid mixing than would ever be possible from molecular motion alone (Cussler, 1997). In particular, for turbulent flow in a pipe, the dispersion coefficient is given by

$$E = dv/2 \quad (28)$$

where d is the pipe diameter and v is the average velocity of the fluid in the pipe. However, if the flow in the pipe is laminar instead of turbulent, the corresponding result is

$$E = \frac{d^2 v^2}{192D} \quad (29)$$

Thus in turbulent flow, the dispersion coefficient is independent of the diffusion coefficient, but in laminar flow, the dispersion coefficient depends inversely on the diffusion coefficient. This counterintuitive inverse dependence, the result of axial convection coupled with radial diffusion, is the foundation of the Goulay equation describing peak spreading in chromatography. We now return from this dispersion tangent back to diffusion and in particular, to mass transfer.

III. MASS TRANSFER

We now turn to a completely different method of describing diffusion, one that has its greatest value in industrial situations. It is related to both diffusion and dispersion but has a simpler mathematical description. This means that it's more approximate. Unfortunately, it's complicated by questions of units and definitions, which give it a reputation of being a difficult subject.

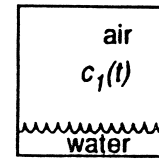
To understand mass transfer, imagine that we have a small amount of water in a large box like that shown in Fig. 5a. The air in the box is originally dry. We want to describe the water concentration in the box—the humidity—as a function of time. Again, we begin with a mass balance like the following

$$\text{accumulation} = [\text{flow in} - \text{out}] + \text{evaporation} \quad (30)$$

Because there is no flow in or out of the box, those terms are zero and the mass balance simply becomes

$$V \frac{dc_1}{dt} = AN_1 \quad (31)$$

(a) Humidification



(b) Packed Bed

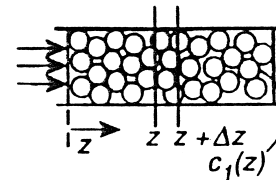


FIGURE 5 Two easy mass transfer examples. In the unsteady case in (a), the water evaporates into the air. In the steady-state case in (b), the spheres are always wet with water, which again evaporates.

where c_1 is the concentration of water vapor in the volume V of the box, A is the surface area of the water, and N_1 is the interfacial flux of the evaporating water. The idea that the total amount of water which evaporates is proportional to the area is straightforward: after all, that's why we spread out rain drops on a tennis court in order to dry the tennis court faster.

The flux N_1 is closely related to the flux j_1 used in the diffusion section (Cussler, 1997; Taylor *et al.*, 1993). The flux here differs because it potentially includes both diffusion and diffusion-induced convection, a distinction which is unimportant when the solute is dilute. We will discuss only that case here. We also will assume that the flux at the interface N_1 is given by

$$N_1 = k(c_{1(sat)} - c_1) \quad (32)$$

where $c_{1(sat)}$ is the water concentration at the interface, which is at saturation. If the air is initially dry, we can combine Eqs. (31) and (32) and integrate to find

$$\frac{c_1}{c_{10}} = 1 - e^{-kat} \quad (33)$$

where $a (= A/V)$ is the liquid area per system volume and k is a new rate constant, called unpoetically a mass transfer coefficient. This simple exponential is the most common result of analysis of mass transfer.

Similar relationships can be developed for steady-state mass transfer. For example, imagine that we have dry air flowing evenly through a bed of wet spheres, like those

shown in Fig. 5b. The concentration of water in the exiting air c_1 will be given by (Cussler, 1997)

$$\frac{c_1}{c_{1(sat)}} = 1 - e^{-ka(z/v)} \quad (34)$$

where z is now the distance from the entry of the bed and v is the velocity of air flowing through the bed. This equation is essentially equivalent to the previous one, but with the residence time (z/v) replacing the actual physical time. Again, it suggests a way in which we can organize data using a mass transfer coefficient k .

But what exactly is being done? We are replacing our detailed description of diffusion of the water with a much more approximate analysis. We are assuming that the bulk of the air is mixed enough to give it a constant concentration. We are assuming that the only significant concentration change occurs close to the water/air interface. This type of analysis and the equations it implies treat

mass transfer like a first-order chemical reaction, but a reversible reaction with an equilibrium constant of one. The equilibrium constant equals one because diffusion is the same in both directions. Nonetheless, the mass transfer coefficient is unlike a chemical reaction because it does not describe chemical change. It describes changes with position or time.

A. Mass Transfer Coefficients

Experimental values of mass transfer coefficients can be collected as dimensionless correlations. One collection of these correlations is in Table II (Cussler, 1997). Because heat transfer is mathematically so similar to mass transfer, many assert that other correlations can be found by adapting results from the heat transfer literature. While this is sometimes true, the analogy is frequently overstated because mass transfer coefficients normally apply across

TABLE II Useful Correlations of Mass Transfer Coefficients for Fluid–Fluid Interfaces

Physical situation	Basic equation ^b	Key variables	Remarks
Liquid in a packed tower	$k \left(\frac{1}{v_g} \right)^{1/3} = 0.0051 \left(\frac{v^0}{av} \right)^{0.67} \left(\frac{D}{v} \right)^{0.50} (ad)^{0.4}$ $\frac{kd}{D} = 25 \left(\frac{dv^0}{v} \right)^{0.45} \left(\frac{v}{D} \right)^{0.5}$ $\frac{k}{v^0} = \alpha \left(\frac{dv^0}{v} \right)^{-0.3} \left(\frac{D}{v} \right)^{0.5}$	a = Packing area per bed volume d = Nominal packing size d = Nominal packing size d = Nominal packing size	Probably the best available correlation for liquids; tends to give lower values than other correlations The classical result, widely quoted; probably less successful than above Based on older measurements of height of transfer units (HTUs); α is of order one
Gas in a packed tower	$\frac{k}{aD} = 3.6 \left(\frac{v^0}{av} \right)^{0.70} \left(\frac{v}{D} \right)^{1/3} (ad)^{-2.0}$ $\frac{kd}{D} = 1.2(1 - \epsilon)^{0.36} \left(\frac{dv^0}{v} \right)^{0.64} \left(\frac{v}{D} \right)^{1/3}$	a = Packing area per bed volume d = Nominal packing size d = Nominal packing size ϵ = Bed void fraction	Probably the best available correlation for gases Again, the most widely quoted classical result
Pure gas bubbles in a stirred tank	$\frac{kd}{D} = 0.13 \left(\frac{P/V}{\rho v^3} d^4 \right)^{1/4} \left(\frac{v}{D} \right)^{1/3}$	d = Bubble diameter P/V = Stirrer power per volume	Note that k does not depend on bubble size
Pure gas bubbles in an unstirred liquid	$\frac{kd}{D} = 0.31 \left(\frac{d^3 g \Delta \rho / \rho}{v^2} \right)^{1/3} \left(\frac{v}{D} \right)^{1/3}$	d = Bubble diameter $\Delta \rho$ = Density difference between gas and liquid	For small swarms of bubbles rising in a liquid
Large liquid drops rising in unstirred solution	$\frac{kd}{D} = 0.42 \left(\frac{d^3 \Delta \rho g}{\rho v^2} \right)^{1/3} \left(\frac{v}{D} \right)^{0.5}$	d = Bubble diameter $\Delta \rho$ = Density difference between bubbles and surrounding fluid	Drops 0.3-cm diameter or larger
Small liquid drops rising in unstirred solution	$\frac{kd}{D} = 1.13 \left(\frac{dv^0}{D} \right)^{0.8}$	d = Drop diameter v^0 = Drop velocity	These small drops behave like rigid spheres
Falling films	$\frac{kz}{D} = 0.69 \left(\frac{zv^0}{D} \right)^{0.5}$	z = Position along film v^0 = Average film velocity	Frequently embroidered and embellished

Notes: ^a The symbols used include the following: D is the diffusion coefficient; g is the acceleration due to gravity; k is the local mass transfer coefficient; v^0 is the superficial fluid velocity; and v is the kinematic viscosity.

^b Dimensionless groups are as follows: dv/v and v/av are Reynolds numbers; v/D is the Schmidt number; $d^3 g (\Delta \rho / \rho) v^2$ is the *Grashoff number*. kd/D is the *Sherwood number*; and $k/(vg)^{1/3}$ is an unusual form of *Stanton number*.

fluid-fluid interfaces. They describe mass transfer from a liquid to a gas or from one liquid to another liquid. Heat transfer coefficients normally describe transport from a solid to a fluid. This makes the analogy between heat and mass transfer less useful than it might at first seem.

The correlations in Table II are most often written in dimensionless numbers. The mass transfer coefficient k , which most frequently has dimensions of velocity, is incorporated into a Sherwood number Sh

$$Sh = \frac{kd}{D} \quad (35)$$

where d is some characteristic length, like a pipe diameter or a film thickness, and D is the same diffusion coefficient which we talked about earlier. The mass transfer coefficient is most frequently correlated as a function of velocity, which often appears in a Reynolds number Re

$$Re = \frac{dv}{\nu} \quad (36)$$

where v is the fluid velocity and ν is the kinematic viscosity; in a Stanton number St

$$St = \frac{k}{v} \quad (37)$$

or as a Peclet number Pe

$$Pe = \frac{dv}{D} \quad (38)$$

The variation of mass transfer coefficients with other parameters, including the diffusion coefficient, is often not well studied, so the correlations may have a weaker experimental basis than their frequent citations would suggest.

B. Problems with Mass Transfer Coefficients

Mass transfer coefficients are frequently regarded as a difficult subject, not because the subject is inherently difficult, but because of different definitions and because of complexities for mass transfer from one solution into a second solution. These differences merit further discussion.

The complexities of definitions occur primarily because concentration can be expressed in so many different variables. In the above, we have assumed that it is expressed in mass per volume or moles per volume. The concentration can equally be well expressed as a mole fraction, which in the liquid phase is commonly indicated by the symbol x_1 and in a gas phase is written as y_1 . In gases, one can also express concentrations as partial pressures. In some cases, especially in medicine, the concentration can be expressed in other more arcane units. For example, “oxygen tension” measures the amount of oxygen present in blood, but it is expressed as the partial pressure that would exist

TABLE III Common Forms of Mass Transfer Coefficients

Basic equation ^a	Typical units of k^b	Remarks
$N_1 = k \Delta c_1$	cm/sec	Common in the older literature; used here because of its simple physical significance
$N_1 = k_p \Delta p_1$	mol/cm ² -sec – atm	Common for a gas absorption; equivalent forms occur in biological problems
$N_1 = k_x \Delta x_1$	mol/cm ² -sec	Preferred for practical calculations, especially in gases
$N_i = k \Delta c_1 + c_1 v^0$	cm/sec	Used in an effort to include diffusion-induced convection

Notes: ^a In this table, N_1 is defined as moles/L²t, and c_1 as moles/L³. Parallel definitions where N_1 is in terms of M/L^2t and c_1 is M/L^3t are easily developed. Definitions mixing moles and mass are infrequent.

^b For a gas of constant molar concentration c , $k = RTk_p = k_y/c$. For a dilute liquid solution $k = (M_2/\rho)k_x$, where M_2 is the molecular weight of the solvent, and ρ is the solution density.

in a gas phase which was an equilibrium with the blood at the experimental conditions.

Each of these units of concentration may be used to define a different mass transfer coefficient, as exemplified by the definitions in Table III. It is not a difficult task to convert a value from one form of coefficient into another form of coefficient (Cussler, 1997; Treybal, 1980). However, it is complicated and requires care. It’s like balancing a check book: it doesn’t always work out the first time you try it. Still, we normally find that with the definitions like those in Table III held firmly in mind, we can readily convert from one form of coefficient to another.

The second reason that mass transfer coefficients are considered difficult happens when mass transfer occurs from one fluid phase into another. This is a genuine source of difficulty, where confusion is common. To see why the difficulty occurs, imagine we are extracting bromine from water into benzene. When we begin, the bromine is at a higher concentration in the water than in the benzene (Cussler, 1997). Later on, the concentrations in water and benzene become equal. Still later, the concentration in the water will have dropped well below that in the benzene. Even then, bromine can still be diffusing from its low concentration in the water into its much higher concentration in the benzene.

The reason that this occurs is that bromine is much more soluble in benzene than it is in water. It partitions from water into benzene. At equilibrium, the concentration in benzene divided by that in water will be a constant much greater than one, and almost independent of the initial concentration of the bromine in the water. Phrased in other terms, in the eventual equilibrium, the concentrations are

not equal. The free energies are equal, but free energy is a considerably more difficult concept than concentration.

The result of this chemistry is that the mass flux across an interface from one phase into the other is not directly proportional to the concentration difference between the two phases. Instead, it is proportional to the concentration in the one phase minus the concentration that would exist in the other phase if it were in equilibrium. In the example just given, this concentration difference is the value in water minus the value in hypothetical water in equilibrium with benzene. This concentration difference makes the study of mass transfer coefficients difficult.

To make these ideas more quantitative, imagine that we are absorbing sulfur dioxide from a flue gas stream into an aqueous stream. The flux of sulfur dioxide is given by the equations

$$N_1 = k_p(p_1 - p_{1i}) \quad (39)$$

where k_p is the form of mass transfer coefficients based on partial pressure differences, p_1 is the partial pressure of the SO_2 in the bulk gas, and p_{1i} is the partial pressure in the gas at the gas/liquid interface. This flux is also given by

$$N_1 = k_x(x_{1i} - x_1) \quad (40)$$

where x_{1i} is the mole fraction of SO_2 at the gas/liquid interface but in the liquid, and x_1 is the mole fraction of SO_2 in the bulk liquid. While these interfacial concentrations are almost always unknown, they are related by a Henry's law constant H :

$$p_{1i} = Hx_{1i} \quad (41)$$

When we combine Eqs. (35)–(37), we obtain the relationship

$$N_1 = \left[\frac{1}{\frac{1}{k_p} + \frac{H}{k_x}} \right] (p_1 - Hx_1) \quad (42)$$

This result is frequently written as

$$N_1 = K_p(p_1 - p_1^*) \quad (43)$$

where the overall mass transfer coefficient K_p is equal to the quantity in square brackets in Eq. (42) and the hypothetical partial pressure p_1^* is simply equal to Hx_1 . This p_1^* is the partial pressure that would exist in the gas if the gas were in equilibrium with the liquid.

This analysis is difficult, and takes careful thought to understand. The key test is to constantly ask what happens at equilibrium. At equilibrium, the partial pressure difference, or the mole fraction difference, or the concentration difference must be zero. The only question is does that difference represent an actual concentration or some

sort of fictional concentration difference designed for our convenience.

IV. CONCLUSIONS

Diffusion, dispersion, and mass transfer are three ways to describe molecular mixing. Diffusion, the result of molecular motions, is the most fundamental, and leads to predictions of concentration as a function of position and time. Dispersion can follow the same mathematics used for diffusion, but it is due not to molecular motion but to flow. Mass transfer, the description of greatest value to the chemical industry, commonly involves solutes moving across interfaces, most commonly, fluid-fluid interfaces. Together, these three methods of analysis are important tools for chemical engineering.

NOTATION

a	Surface area per volume
A	Area
c_1	Concentration of species "1"
d	Pipe diameter
D	Diffusion coefficient
D_{ij}	Diffusion coefficients in multicomponent systems
E	Dispersion coefficient
H	Henry's law constant
j_1	Diffusion flux of species "1"
k, k_p, k_x	Mass transfer coefficients
k_B	Boltzman's constant
K_p	Overall mass transfer coefficient
l	Length or thickness
M	Total solute mass in pulse
n_1	Total flux of species "1"
N_1	Interfacial flux of species "1"
p	Total pressure
p_1	Partial pressure of species "1"
S	Amount solute emitted per time
T	Temperature
t	Time
v_1, v^0	Velocity of species "1" and of reference, respectively
V	Volume
x	Velocity direction
x_1, y_1	Mole fractions of species "1" in liquid and gas, respectively
z	Position
γ_1	Activity coefficient of species "1"
μ	Viscosity
μ_1	Chemical potential

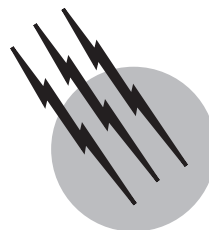
SEE ALSO THE FOLLOWING ARTICLES

FLUID DYNAMICS • FLUID MIXING • HEAT TRANSFER
• LIQUIDS, STRUCTURE AND DYNAMICS • MOLECULAR
HYDRODYNAMICS • PLASTICIZERS

BIBLIOGRAPHY

- Carslaw, H. S., and Jaeger, J. C. (1986). "The Conduction of Heat in Solids," 2nd ed., Clarendon, Oxford.
- Crank, J. (1975). "The Mathematics of Diffusion," 2nd ed., Clarendon, Oxford.
- Cussler, E. L. (1977). "Diffusion," 2nd ed., Cambridge University Press, Cambridge.

- de Groot, S. R., and Mazur, P. (1962). "Non-Equilibrium Thermodynamics," North-Holland, Amsterdam.
- Fick, A. E. (1855). *Poggendorff's Ann. Phys.* **94**, 59.
- Graham, T. (1850). *Phil. Trans. R. Soc.* **140**, 1.
- Katchalsky, A., and Curran, P. F. (1967). "Non-Equilibrium Thermodynamics in Biophysics," Harvard University Press, Cambridge.
- Kim, S., Kohl, M., and Myerson, A. S. (1997). *J. Crystal Growth* **181**, 61.
- Reid, R. C., Sherwood, T. K., and Prausnitz, J. M. (1977). "Properties of Gases and Liquids," 3rd ed., McGraw-Hill, New York.
- Seinfeld, J. H. (1985). "Atmospheric Chemistry and Physics of Air Pollution," Wiley, New York.
- Taylor, R., and Krishna, R. (1993). "Multicomponent Mass Transfer," Wiley-Interscience, New York.
- Treybal, R. E. (1980). "Mass Transfer Operations," McGraw-Hill, New York.



Membranes, Synthetic, Applications

Eric K. Lee

Integrated Biosystems Inc.

W. J. Koros

Georgia Institute of Technology

- I. General Principles
- II. Membrane Materials, Geometry, and Packaging
- III. Gas Separations
- IV. Vapor–Liquid Separations
- V. Liquid Separations
- VI. Biotechnology and Life Sciences
- VII. Biomedical Applications
- VIII. Membrane Sensors

GLOSSARY

Membrane Structure, having lateral dimensions much greater than its thickness, through which mass transfer may occur under a variety of driving forces.

Asymmetric membrane Membrane constituted of two or more structural planes of non-identical morphologies.

Composite membrane Membrane having chemically or structurally distinct layers.

Homogeneous membrane Membrane with essentially the same structural and transport properties throughout its thickness.

Synthetic (artificial) membrane Membrane formed by a process not occurring in nature.

Upstream Side of a membrane into which penetrants enter from the feed stream.

Stage cut Parameter defined as the fractional amount of the total feed entering a membrane module that passes through the membrane as permeate.

Penetrant (permeant) Entity from a phase in contact with one of the membrane surfaces that passes through the membrane.

Membrane module (cell) Manifold assembly containing a membrane or membranes to separate the streams of feed, permeate, and retentate.

Membrane reactor Device for simultaneously carrying out a reaction and membrane-based separation in the same physical enclosure.

SYNTHETIC MEMBRANES¹ are thin barriers that allow preferential passage of substances on a microscopic or molecular size level. Starting with this single attribute, a broad area of science and technology has evolved over the past century where membrane processes are used as efficient and economical methods of separation and purification. Today, membrane processes contribute to many sectors of scientific research and development, industry, medicine, and management of natural and man-made resources. Many membrane applications are so deceptively simple that the physical science governing their use is easily overlooked. The field is best partitioned into smaller topical areas to understand the diverse types and uses that membranes have in nature and industry. The present article is organized according to this systematic approach.

A membrane, whether naturally occurring or synthetic, is taken to be a structure with a large aspect ratio in which one of its three dimensions is much thinner than the other two dimensions. The simplest form of a membrane is thus a flat diaphragm, but the above description also applies to hollow fiber, or even a spherical or bag-like encapsulation domain surrounding living cells.

I. GENERAL PRINCIPLES

The discussion of synthetic membranes can be structured in terms of the “function” or the “structure” of the membrane used in a particular application. For instance, one can consider whether a membrane is used to separate mixtures of gas molecules vs particles from liquids (function) vs whether the membrane structure is primarily microporous or dense (structure). In fact, function and structure are linked, but to facilitate the consideration of physical science issues related to membranes appropriate for this reference, emphasis on functional aspects are probably most appropriate. This approach reflects the fact that the use of a membrane generally involves one or more physical sci-

ence principles, such as diffusion or fluid flow. By understanding the principles controlling function, the required structure to enable that function becomes clear. This will be illustrated for several examples, and the broader topic of additional physical science phenomena that are potentially useful in future or emerging membrane processes will also be noted, even if practical commercial examples may not yet exist.

In use, most synthetic membranes involve a transport of one or more components from an “upstream” side of the membrane to a “downstream” side. Although microscopic interpretations differ between the various applications, description of the transport process for a component, A, from the upstream to the downstream side of the membrane is possible in terms of Eq. (1):

$$n_A = [(Driving\ Force)_A] / (Resistance)_A = [(DF)_A] / \Omega_A, \quad (1)$$

where n_A is the flux of A, equal to the rate of transfer of component A per unit area per unit time. The net driving force $(DF)_A$ acting on component A between the upstream and downstream membrane face and the net resistance retarding movement of A (Ω_A) , while simple to write, may have complex physical chemical origins that differ greatly between the various types of membrane applications. Despite these limitations, Eq. (1) is useful to unite the discussion, since it provides a framework to understand the essential nature of most membranes.

One can devise an almost unlimited number of net driving force terms, DF_A , by imposing a difference in any intensive thermodynamic variable between the upstream and downstream membrane faces. Coupling between the effects can occur, but generally one driving force, e.g., pressure, temperature, concentration, or voltage, is sufficiently dominant in a given application to allow focusing on it primarily.

The resistance term in Eq. (1), Ω_A , usually increases directly with the membrane thickness, so reducing thickness by some percentage generally increases flux by the same percentage. This generalization has some exceptions. For instance, reaction or complexation kinetics within the membrane or nonhomogeneous morphologies within the membrane can cause such exceptions in some cases (Crank, 1975).

A. Major Membrane Application Types

To facilitate the discussion, conventional terminology used to refer to the most common types of membrane-based processes is presented in Table I along with typical driving forces used in each application.

¹The most obvious division of the membrane world occurs between synthetic (man-made) and biological (naturally occurring) materials. The present discussion will focus only on *synthetic membranes*, which alone is an enormous area. Biological membranes have been the topics of books and reviews (Yeagle, 1992) at least as extensive as that of synthetic membranes. Despite sharing interest in the large aspect ratio nature common to all membranes, the two fields have developed quite separately. In any case, the physical science related to synthetic membranes is fairly well understood and provides a useful basis for understanding many aspects of the more complex biological membrane topical area.

TABLE I Primary Synthetic Membrane Applications and Driving Forces

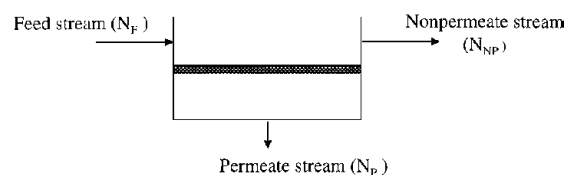
Function or application	Typical driving force type
Membrane dialysis (D)	Concentration
Microfiltration (MF)	Pressure (10–25 psi)
Ultrafiltration (UF)	Pressure (10–100 psi)
Nanofiltration (NF)	Pressure (100–500 psi)
Reverse osmosis (RO)	Pressure (minus osmotic pressure) (100–1500 psi)
Gas separation (GS)	Partial pressure (10–1000 psi)
Pervaporation (PV)	Activity, effective partial pressure
Carrier facilitated transport (CFT)	Activity, concentration
Ion conduction	Ion concentration, voltage
Ion exchange	Electrochemical interactions
Affinity separation	Biospecific interactions

Many controlled release devices are not “membranes” by the conventional definition, since only transient release of an active agent, without permeation occurring between an upstream and a downstream, is typical. Nevertheless, some controlled release units do operate with a concentration driving force to achieve effectively steady state release from the internal reservoir of the device to the external surrounding. Such processes are included here for completeness.

Membrane reactors and contactors for extraction, gas absorption, or membrane distillation represent extensions of various types of the membranes in Table I and Table II. Nevertheless, these cases, along with controlled release of application, will be considered briefly to illustrate how the basic membrane types in Table I can be applied in unconventional, ever-expanding ways.

TABLE II Characteristic Penetrant Size (Diameter) Spectrum for Nonpermeating Species

Application	Nonpermeating species size
Conventional (nonmembrane) filtration	>200,000 Å
Microfiltration (MF)	1,000–200,000 Å
Ultrafiltration (UF)	20–100 Å (MW 10,000–100,000)
Membrane dialysis (D)	5–50 Å (MW 50–10,000 daltons)
Nanofiltration (NF)	5–20 Å
Reverse osmosis (RO)	3–5 Å (hydrated microsolute and ions)
Gas separation (GS)	3–5 Å
Pervaporation (PV)	3–5 Å
Carrier facilitated transport (CFT)	3–10 Å (gases and dissolved solutes)
Ion conduction (IC)	3–5 Å

**FIGURE 1** Idealized membrane process showing feed (N_F), nonpermeate (N_{NP}) and permeate (N_P) streams.

Most membrane operations indicated in Table I are run as continuous steady state processes with a feed, permeate, and retentate stream (see Fig. 1). For example, in dialysis, a feed stream comprising blood with urea and other metabolic by-products passes across the upstream face of a membrane while an electrolyte solution without these by-products passes across the lower face of the membrane. A flux of by-products (A) occurs into the downstream where it is taken away as a permeate and the purified blood leaves as nonpermeate.

In microfiltration and ultrafiltration a feed stream containing suspended particles passes across the upstream face of a membrane at a higher pressure than exists at the downstream. This pressure driving force motivates the suspending fluid (usually water) to pass through physically observable pores in the membrane. This process achieves a concentration of the particles or macromolecules in the nonpermeate stream and produces essentially pure particle-free permeate. Such processes are extremely useful for processing of thermally labile feeds and are even being used as replacements for sand filters in water clarifying and purification. Cost is generally an important issue, so minimization of the membrane resistance Ω_A in Eq. (1) requires a small effective membrane thickness to achieve high fluxes at low pressure differences. This theme, the need to achieve a very small effective thickness, runs throughout most of the membrane applications, since cost is related to required membrane area and required membrane area is inversely proportional to the achievable flux (Koros, 1995).

In the other pressure-driven separations in Table I, the difference in size between the permeating component A and rejected components B , C , etc., is progressively reduced in NF vs RO vs GS. This shift in size discrimination requirements is illustrated in Table II.

Recently, impressive strides have been made in controlling the effective sizes of suspended macromolecules by adjusting ionic strength and pH to selectively alter the effective size in solution of nominally similar molecular weight components. This approach allows the smaller of the two components to pass through the membrane with the suspending solvent to the permeate to allow fractionation of two similarly sized dissolved macromolecules.

Ultimately, strictly *hydrodynamic sieving* of a suspending solvent A (typically $<3\text{--}4\text{ \AA}$) away from suspended B becomes inadequate and molecular forces enter as dominant factors. Arguments are still heard regarding the need for combined hydrodynamic and molecular factors in NF separations. Nevertheless, the boundary at which molecular factors become dominant is typically set between ultrafiltration and nanofiltration for such pressure-driven cases.

Indeed, by their nature, membranes combine thermodynamically based partitioning and kinetically based mobility discrimination in an integrated separation unit. In membranes with large pores relative to the size of individual molecules, the thermodynamic partitioning aspects are practically negligible in most cases. Except for MF and UF, however, both aspects must be considered to some degree depending upon the specific application. Examples of the limiting cases between these two situations can be seen in the case of separation of O_2 and N_2 using either carrier-facilitated transport or molecular sieving mechanisms for gas separation. In the case of carrier-facilitated transport, thermodynamics dominates the selective separation of one gas from another. In the case of molecular sieving, size and shape dominate, but in a more complex manner related to actual molecular architecture as compared to simple hydrodynamic sieving seen in microfiltration and ultrafiltration (Koros and Mahajan, 2000).

B. Mechanisms of Membrane Separation

1. Hydrodynamic Sieving

a. Simple effects. The sieving mechanism responsible for rejecting large particles, colloids, and even macromolecules from suspending low molecular weight fluids in MF and UF depends upon classical hydrodynamic forces. Conventional filtration experience is a useful starting point for understanding ultrafiltration; however, colloidal and even molecular scale forces play a larger role in the case of micro- and especially ultrafiltration. The suspending solvent can be viewed as a continuum, and its flow through the porous membrane can be described by Darcy's law (Belfort, Davis, and Zydney, 1994).

In the absence of suspended solutes or colloids, the pure solvent flux through an ultrafiltration membrane is directly proportional to the applied pressure difference and inversely proportional to the viscosity of the solvent and the membrane thickness. Transport within the pores occurs in the "creeping flow" regime, since kinematic viscosities of liquids are sufficient to make $\text{Re} \ll 1$ for practical pore sizes. In the simplest case, the membrane can be considered to be a packed array of straight, equal diameter nonintersecting capillary tubes. The observed volumetric flux, $n_A \hat{v}_A$ (cc/sec cm^2), equals the product of the mass flux of solvent based on the total membrane area, n_A

(g/sec cm^2), and the solvent specific volume, $\hat{v}_A$ (cm^3/g). This volumetric flux, relates to the average velocity in a pore, $\bar{U}_{Ap}$ as follows (Carman, 1937):

$$n_A \hat{v}_A = \varepsilon [L/L_e] \bar{U}_{Ap}, \quad (2)$$

where ε is the membrane porosity (pore volume/total volume) and L/L_e is the ratio of the physical thickness of the membrane to the "effective" distance a flowing penetrant must travel due to complex morphology of the medium. For straight cylindrical pores oriented perpendicular to the membrane upstream and downstream surfaces, $L/L_e = 1$, and the expression for flux relates simply to the fluid viscosity, η , pore diameter, d_p , and pressure drop across the membrane, Δp , according to Eq. (3) (Cheryan, 1986):

$$(n_A \hat{v}_A)_o = [\varepsilon d_p^2 / 32 \eta L] \Delta p. \quad (3)$$

The subscript "o" indicates that this flux is for solvent "A" without the presence of rejected solute "B." This model is appropriate for membranes that have straight pores; however, such membranes are not typical of most industrial membranes. Not only are the rate limiting pores in the actual working skin layer tortuous, but a complex porous support layer is generally present that supports the working skin layer. In well-made sieving membranes, the porous support is ideally "invisible" to the transport process, as is illustrated in Fig. 2. The support, therefore, simply serves as a scaffold for the ideal selective layer. Formation of such structures requires some care, but technology exists to achieve this requirement as is described elsewhere (Koros and Pinnau, 1994).

Even if the selective layer comprises a well-defined thin region, L , atop a porous support, the pores through this region are not typically straight. A more general expression, the Carman-Kozeny equation, is useful for complex skin morphologies in membranes formed by various precipitation processes (Leenaars and Burggraaf, 1985). This expression is based on the assumption that the liquid flows through individual tortuous paths with a volume-to-surface ratio equal to the experimentally measured pore volume divided by the internal surface area of the selective layer. The pores need not be straight, but all pores are assumed to have the same effective diameter. Techniques for pore volume and surface area measurements are well known, but the nonhomogeneous or "asymmetric" nature of typical membranes makes it complex to characterize the skin layer independently of the support. When this



FIGURE 2 Idealized high-performance synthetic membrane showing a top rate-limiting selective layer and an "invisible" (with regard to transport) support layer.

can be done, a commonly used characteristic of pores is their hydraulic diameter defined in Eq. (4).

$$d_h = 4[\text{total pore volume/total internal pore area}] \quad (4)$$

The Carman–Kozeny equation for flux through such a porous medium is

$$(n_A \hat{v}_A)_o = [\varepsilon d_h^2 / (16\eta k_o k_t L)] \Delta p = k_h \Delta p, \quad (5)$$

where k_o is a shape factor that accounts for the various possible pore shapes that result from packing of different shaped particles comprising the selective layer. The complex flow path length is accounted for by introducing another parameter, $k_t = (L_e/L)^2$, which is usually called the “tortuosity.” Note, however, that the term “tortuosity” is also used by some authors to equal L_e/L , so care must be used when referring to the term tortuosity. The term, k_h , is referred to as the hydraulic permeability (Coulson, 1949). Clearly, in this case, if Δp is taken as the driving force term in Eq. (1), the resistance term in Eq. (1) is a complex function of the selective layer of the membrane and the fluid properties. For packed beds of particles resulting in effective pathways of average length, L_e , the so-called Kozeny constant, $k_o k_t$, is near 5 (Carman, 1937; Leenaars and Burggraaf, 1985; Coulson, 1949). Accepting this value allows one to determine k_h from pure solvent flux vs pressure drop. For complex membrane morphologies, such an idealization is theoretically questionable. Nevertheless, this approach provides a useful characterization parameter for the effective diameter of the pores in the thin region of the membrane that controls flux. To achieve high fluxes, membranes are often asymmetric with a thin selective region supported on an open-cell support material. A flow-based characterization of k_h , is especially valuable for these complex media. Such characterizations avoid the need to measure the detailed values from Eq. (5) in the thin flux-determining layer at the membrane surface, but it should be remembered that it is a highly simplistic view of reality. Despite its limitations, the Carman–Kozeny equation remains popular, since more realistic descriptions of practical membranes are encumbered by the lack of ability to characterize the actual detailed morphology of the membrane.

The volumetric flux of a binary mixture of solvent and solute, j_v , through a membrane can be expressed as

$$j_v = n_1 \hat{v}_A + n_2 \hat{v}_B, \quad (6)$$

where $n_A \hat{v}_A$ is the solvent flux in the presence of the solute and n_B and $\hat{v}_B$ are the mass flux and partial specific volume of the solute, respectively. For dilute solute concentrations $j_v \sim n_A \hat{v}_A$, and under dilute ideal conditions, $j_v \sim n_A \hat{v}_A \sim (n_A \hat{v}_A)_o$ [see Eq. (5)] (Miller, 1992). Complications due to so-called “concentration polarization” and “fouling” often invalidate this simple approximation,

as described later. First, however, an additional issue related to complex flow fields *within* the pores of the membrane should be considered. When the dimensions of the pores and those of the suspended solute are similar, such as in the case of nanofiltration (NF) a process called “hindered transport” occurs (Deen, 1987).

b. Complex effects.

i. Hindered transport. Intuition correctly predicts complete rejection of a monodisperse solute from a solvent “A” by a membrane with a uniform pore size *smaller* than the solute “B” but much larger than the solvent “A.” Rejection “R” of component “2” is defined as $R = 1 - C_{\text{downstream}}/C_{\text{upstream}}$, where the concentrations refer to solute B downstream and upstream, respectively. Again, in this case solvent flux can be described in terms of Eq. (5).

Surprisingly, intuition *fails* to predict the behavior of the same solute and solvent in a membrane with a uniform pore size larger than *both* the solvent and solute. The expectation that such a membrane will provide no rejection of the solute has been refuted repeatedly. Indeed, careful experiments indicate that partial rejection of the solute occurs *even* when the solute is considerably smaller (say 1/10th as large as the pore size) (Miller, 1992; Deen, 1987; Ho and Sirkar, 1992; Happel and Brenner, 1965). The extent of rejection increases monotonically to the total rejection limit as the solute size approaches the pore size. These effects arise both from entropic suppression of partitioning *and* from augmented hydrodynamic resistance to transport through the fine pores. Thus, in this case, for a porous membrane, thermodynamic partitioning can play a role in the physical chemical processes of transport.

For a solute of finite dimensions, a decrease in solute entropy occurs upon partitioning from an unbounded external solution into a confined pore space. The decrease in entropy results in a lower solute concentration in the pore compared to the external solution to allow equalization of chemical potential with the solute in the external fluid. For cylindrical pores, $d_h = d_p$, and the partition coefficient, $K_i = (C_i)_{\text{internal}}/(C_i)_{\text{external}}$, between the internal and external solutions is $K_i = (1 - \lambda)^2$, where $\lambda = d_s/d_p$ the ratio of solute to pore diameter (Happel and Brenner, 1965). Even for solutes 50% as large as the pore diameter, this factor equals 0.25, yielding a fourfold reduction in concentration within the pore. As λ approaches unity, K_i drops tremendously. Related expressions exist for other geometries, and the trends are similar (Happel and Brenner, 1965).

Hindered transport of a solute moving within a continuum of solvent in a small pore can be analyzed in terms of classical hydrodynamics (Deen, 1987). The penetrant-to-pore size ratio (λ) and the position of the penetrant within a

pore allow calculation of transport hindrance of the solute based on properties of the solvent. For dilute solute concentrations, the solvent flows with a well-defined velocity profile and average velocity that is characteristic of the pore cross-sectional shape. The solute, on the other hand, moves with its own characteristic velocity relative to the pore wall and is assumed to sample all radial positions available to it during its transport. An expression relating the local steady state solute flux to a diffusive driving force *and* a convective bulk flow term is averaged across the pore cross section. This procedure yields the steady state solute flux in terms of the cross sectionally averaged solute concentration and solvent velocity along the length of the pore.

Effectively, the theory results in the addition of two cross-sectionally averaged intrapore hydrodynamic hindrance coefficients for diffusion and convection, respectively. Solute hindrance, and hence rejection, relative to the solvent, increases qualitatively with increasing total flux. The asymptotic value at high total flux values is determined by the value of λ (Ho and Sirkar, 1992; Mitchell and Deen, 1986; Dalvie and Baltus, 1992). In the absence of hindering effects, both of the intrapore hydrodynamic coefficients equal unity and the solute behaves as it would in an unbounded medium. These coefficients approach unity as the ratio of the solute diameter to the pore diameter becomes negligible, i.e., $\lambda \rightarrow 0$. Moreover, since the entropic partitioning effect also disappears in this limit, no rejection occurs as the ratio of the solute diameter to the pore diameter becomes negligible. The detailed application of this elegant model is often compromised by *distribution of pore sizes in practical membranes*. For a given penetrant, a pore size distribution translates to a distribution in λ , thereby complicating applications in cases of nonisoporous media comprising the majority of practical membranes. Characterization of the wide range of pore distributions in membranes seen in practical membranes requires several complementary techniques. Useful reviews of ultrafiltration, microfiltration, and dialysis membranes, and characterizations of their pore size distributions deal with this topic in detail (Belfort, Davis, and Zydney, 1994; Sakai, 1994).

As noted above, as the size difference between the solvent and solute become progressively smaller, viscous flow rapidly becomes less important, and molecular interactions become dominant factors. In this limit, molecular solution (or sorption) and diffusion phenomena control the relative transport rates of the solute and solvent. This transition region is an area of ongoing discussion regarding “what is a pore and what is not a pore?”

ii. *Concentration polarization and fouling in MF, UF, and NF.* Concentration polarization typically refers to a reversible buildup of rejected nonpermeated solute in

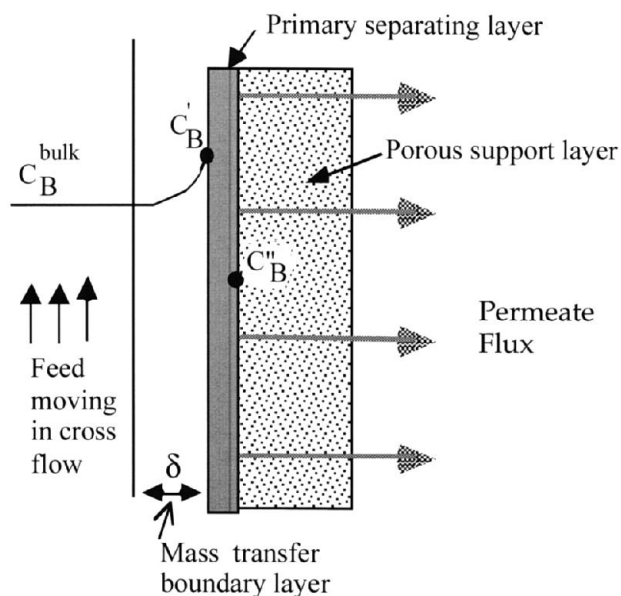


FIGURE 3 Illustration of the concept of an external mass transfer boundary layer resistance and associated concentrations of the rejected species, B.

the external phase near the membrane-solution interface. Solvent convects solute to the membrane face where it is rejected as solvent passes through to produce a permeate stream. This phenomenon is illustrated in Fig. 3, where a mass transfer boundary layer provides resistance to the back diffusion of rejected solute to the well-mixed bulk. At steady state, the flux of solute approaching the membrane equals the flux of unrejected solute leaving in the permeate plus the rejected solute flux arising from back diffusion through the boundary layer (Belfort, Davis, and Zydney, 1994; Cheryan, 1986; Glasstone, 1950). For such a case, the combined solvent-solute permeate volumetric flux can be written as follows:

$$j_v = k_c \ln \left(\frac{C_B' - C_B''}{C_B^{\text{bulk}} - C_B''} \right), \quad (7)$$

where k_c is equal to the solute mass transfer coefficient, and C_B^{bulk} , C_B' , and C_B'' are the solute concentration (g/cm^3) at the local upstream bulk, upstream membrane, and downstream membrane faces, respectively. Viewed in the simplest possible terms, k_c equals the ratio of the Brownian diffusion coefficient (D_{eB}) and the boundary layer thickness (δ) in the external fluid contacting the membrane, i.e., $k_c = D_{eB}/\delta$. Equation (7) indicates that a higher solute concentration exists at the upstream membrane face compared to the bulk for sieving type separations.

Correlations for k_c in both the laminar and turbulent regimes have been found satisfactory for ultrafiltration engineering designs (Cheryan, 1986; Glasstone,

1950; Bessarabov, 1999). With adequate crossflow past the membrane face, the boundary layer thickness is reduced greatly, so for dilute solutions j_v approaches $(n_A \bar{v}_A)_0$, the pure solvent flux. Increasing feed pressure increases j_v and C'_B , which can cause the solute to form a precipitated cake or gel at the membrane face. The C'_B value tends to be a constant characteristic of the precipitated layer, and further pressure increases produce no increase in flux. Excessive feed pressures may not only fail to provide additional flux, but also may complicate subsequent cleaning due to serious occlusion of pores, so operating at pressures leading to limiting flux behavior is generally not recommended.

Qualitatively similar polarization responses are apparent for both microfiltration and ultrafiltration processes, but for microfiltration, additional more complex can occur as is discussed below (Belfort, Davis, and Zydney, 1994; Segre and Silberberg, 1962). The “rejection” parameter mentioned earlier can be written as either an “observed,” R_o , or an “intrinsic,” R_i , value:

$$R_o = 1 - \frac{C''_B}{C'_B} = 1 - \frac{n_B}{C'_B j_v} \quad (8a)$$

and

$$R_i = 1 - \frac{C''_B}{C'_B} = 1 - \frac{n_B}{C'_B j_v}. \quad (8b)$$

The observed rejection [Eq. (8a)] is clearly the important one for a practical separation operations, but it includes the confounding effects of concentration polarization. Since $C''_B \leq C'_B$, the observed rejection is less than the intrinsic rejection and can be determined by estimating the solute wall concentration with Eq. (7).

As the name implies, R_i characterizes the intrinsic ability of the membrane to reject the solute. Molecular weight and size correlate roughly for high molecular weight solutes, and although imprecise, it is common to characterize the intrinsic “cutoff molecular weight” of a given membrane. A membrane can be “calibrated” by determining the molecular weight of a series of standard solutes at which roughly 90% rejection occurs under conditions with negligible concentration polarization. Monodisperse solutes such as polyethylene glycol (PEG) of known molecular weight are useful for this purpose. Table III gives typical values for PEGs in water at 25°C (Segre and Silberberg, 1962; Miller, 1992): The effective size of a macromolecule depends upon the quality of the solvent used, so in a non-aqueous solution, reevaluation of the effective size of the PEG's would be necessary (Segre and Silberberg, 1962; Miller, 1992).

In microfiltration, especially for larger particles (>1–20 μm typically), molecular diffusional phenomena have little impact. In these cases, deposition on the membrane surface is prevented primarily by exploiting various fluid

TABLE III Relationship between Molecular Weight and Hydrodynamic Diameter Estimated Using Intrinsic Viscosity Measurements at 25°C for Essentially Monodisperse PEGs

PEG sample	M_w (g/mole)	d_s (Å)
400	376	12.2
1,000	1,025	19.6
1,500	1,569	24.2
2,000	2,052	27.8
3,000	2,971	33.8
4,000	3,872	38.8
6,000	6,375	51.0
12,000	12,000	81.0
35,000	35,000	133.4

dynamical effects. Early modules maximized membrane packing density without much attention to fluid dynamics, and suboptimal performance resulted (Belfort, Davis, and Zydney, 1994). With suspensions, shell-fed hollow fiber and even spiral wound modules have a tendency to clog, while flat sheet and tubular designs show the least tendency to clog under crossflow filtration. Turbulent cross-flow velocities are required to avoid serious polarization and fouling with domestic wastewaters and cell culture media that tend to form compressible cakes that complicate operation.

Early predictions of suppression of fouling were based only on Brownian back-diffusion of large colloids and particles. These predictions failed to account for additional factors opposing particle deposition. New phenomena were suspected when experimental data showed that flux increased with increasing suspension particle size, rather than showing a greater tendency for cake deposition as expected. Moreover, the flux increased with shear rate to a higher power than one third, which was predicted for molecular diffusion-dominated boundary layers in the traditional Leveque solution (Belfort, Davis, and Zydney, 1994). Several factors explain this “flux paradox” for particles >0.5 μm diameter, including (i) shear-induced diffusion, (ii) inertial lift, and (iii) surface transport. These mechanisms are described in detail in a recent review on crossflow microfiltration (Belfort, Davis, and Zydney, 1994) and are only summarized here.

The simple form in Eq. (7) can be maintained by replacing the Brownian diffusion coefficient in the expression $k_c = D_{eB}/\delta$ by the shear-induced hydrodynamic diffusion coefficient for the particles, D_s . Shear-induced hydrodynamic diffusion of particles is driven by random displacements from the streamlines in a shear flow as the particles interact with each other. For particle volume fractions between 20 and 45%, D_s has been related to

the first power of the shear rate at the membrane surface and the square of the particle size, viz., $D_S = 0.03d_p^2 \gamma_0$. For example, the shear-induced diffusion coefficient for a 1- μm diameter particle at a shear rate of 1000 sec^{-1} is $3 \times 10^{-7} \text{ cm}^2/\text{sec}$ —more than two orders of magnitude higher than for simple Brownian diffusion of such a particle in water at ambient temperature (Belfort, Davis, and Zydney, 1994). Under such conditions, the steady state permeation flux is expected to be proportional to the shear rate and to increase with particle size, consistent with actual data. Shear-induced diffusion is a factor for particles in the range of 0.5–30 μm , which comprises much of the practically important size range for microfiltration (Belfort, Davis, and Zydney, 1994).

The so-called “inertial lift” phenomenon is another factor opposing membrane fouling for microfiltration (Belfort, Davis, and Zydney, 1994). If the conditions are such that the inertial lift velocity is sufficient to offset the opposing permeate velocity, then the particles are not expected to be deposited on the membrane. Inertial lift arises from nonlinear interaction of a particle with the surrounding flow field under conditions where the Reynolds number based on the particle size is large enough to cause the nonlinear inertial terms in the Navier–Stokes equations to be significant (Belfort, Davis, and Zydney, 1994). The inertial lift increases with the cube of the particle size and the square of the tangential shear rate.

Besides the above subtle effects, simple crossflow-induced drag of the deposited cake toward the filter exit can also help prevent excessive cake accumulation. The tangential drag force can be estimated, but the rheology of the cake may be complex, so prediction of this antifouling force is difficult. Nevertheless, maximizing these velocities is useful, since all the above fluid dynamic effects help prevent fouling under high crossflow conditions. Such antifouling measures come at an expense of mechanical energy input in the form of pump work, and hence operational costs for the system. Ongoing work seeks to optimize the use of such mechanical energy inputs to reduce solute accumulation. Unsteady and secondary flows can also be used to help prevent boundary layers stabilization even at relatively low Reynolds numbers. Taylor and Dean vortex flows, rough channels, flow reversals, rotating flows, torsional oscillating flows, and even internally moving wipers have been used in extreme cases with pastes, pulps, foods, pulp, and other difficult to process feeds (Belfort, Davis, and Zydney, 1994).

In addition to fluid dynamics, surface modification of the membrane can reduce the attractive forces or even create repulsive ones between potential fouling solutes and the membrane (Belfort, Davis, and Zydney, 1994).

Combined surface modification and management of fluid dynamics at the membrane surface are effective tools for fouling avoidance.

2. Sorption-Diffusion Separation Mechanisms

As the size difference between penetrants decreases, molecular sorption and diffusion phenomena control their relative permeation rates across the ideal rate-limiting layer in Fig. 2. As noted earlier, so-called nano porous media (e.g., pores ~ 1 –2 nanometers or 10–20 Å diameter) are usually felt to exist at this limit. Dialysis, electrodialysis, and nanofiltration processes operate in this complex region to perform a selective sorting of electrolytes and other small molecules under mild concentration or electrical driving forces. Recent reviews of membrane-related aspects of electrodialysis and hemodialysis are available for the interested reader (Baker, Cussler, Eykamp *et al.*, 1991; Nakao, 1994). The greatest difficulty and ambiguity in defining pore sizes occur as pores approach *micromolecular* dimensions on the order of 5–10 Å and less.

Low salt rejection RO membranes (e.g., $R < 0.5$ for NaCl) are sometimes classified as “nanoporous” and allow retention of sugars and large molecules while permeating small electrolytes. In this case, a hindered transport description of the process would be appropriate with the water and nonrejected electrolytes being treated as a single “fluid” and the rejected sugar considered the solute.

Good quality RO membranes can reject >95–99% of the NaCl from aqueous feed streams (Baker, Cussler, Eykamp *et al.*, 1991; Scott, 1981). The morphologies of these membranes are typically asymmetric with a thin highly selective polymer layer on top of an open support structure. Two rather different approaches have been used to describe the transport processes in such membranes: the solution-diffusion (Merten, 1966) and surface force capillary flow model (Matsuura and Sourirajan, 1981). In the solution-diffusion model, the solute moves within the essentially homogeneously solvent swollen polymer matrix. The solute has a mobility that is dependent upon the free volume of the solvent, solute, and polymer. In the capillary pore diffusion model, it is assumed that separation occurs due to surface and fluid transport phenomena within an actual nanopore. The pore surface is seen as promoting preferential sorption of the solvent and repulsion of the solutes. The model envisions a more or less pure solvent layer on the pore walls that is forced through the membrane capillary pores under pressure.

For truly high rejection reverse osmosis membranes, the “solution-diffusion” description of this process is the most popular and probably the most realistic. In this case, the high osmotic pressure difference between the

salt-containing feed and almost salt-free permeate streams, $\Delta\pi$, must be overcome to drive water to the permeate side. The osmotic pressure difference, $\Delta\pi$, between two solutions of different concentration is the pressure difference that exists when there is no difference in chemical potential of water on the two sides of the membranes.

Neglecting convection effects, the solution-diffusion model gives the following expressions for water (1) and salt (2) molar fluxes through a membrane with a selective layer thickness of L and a transmembrane pressure drop Δp (Merten, 1966):

$$J_A = D_A K_A \hat{V}_A [\Delta p - \Delta\pi]/LRT, \quad (9a)$$

$$J_B = D_B K_B \Delta C_B/L, \quad (9b)$$

where D_A and K_A and D_B and K_B are the diffusion coefficient and partition coefficients for water and salt in the membrane, respectively. The partial molar volume of water, $\hat{V}_A$, is generally well approximated by the pure component molar volume. The observed salt rejection coefficient is given in terms of external bulk salt concentrations (moles/cm³) and known fluxes as shown below:

$$R_o = 1 - \frac{C_B''}{C_B^{\text{bulk}}} = 1 - \frac{j_B}{C_B^{\text{bulk}} j_v} \approx 1 - \frac{j_B}{C_B^{\text{bulk}} j_A \hat{V}_A}. \quad (10)$$

Increasing the $(\Delta p - \Delta\pi)$ term in Eq. (9a) clearly increases rejection, since the flux of solvent (water) increases proportionally to this factor, while the flux of salt is essentially independent of it, within the accuracy of the approximations of the model. A typical example of such behavior is shown in Fig. 4 as a function of feed pressure at 25°C for a brackish water feed with low salt concentration (0.5 wt % or 0.16 mol %). As expected based on Eq. (9a), when the applied transmembrane Δp equals the

transmembrane osmotic pressure difference between the feed and product water (~ 50 psi), the extrapolated water flux is essentially zero.

Application of RO for water production is now a well-accepted and economical process even for higher concentration seawater with osmotic pressures of over 300 psi. Malta, for example, has evolved an economical reliable application of this technology to produce 60% of its potable water supply (Lamendolar and Tua, 1995).

Rejections of other ions besides Na^+ and Cl^- are tunable characteristics of the reverse osmosis membranes that depend upon the intrinsic nature of polymer separating layer and how it has been processed. In general, bivalent ions like Ca^{2+} and SO_4^{2-} are more easily rejected than are monovalent ones like Na^+ and Cl^- .

II. MEMBRANE MATERIALS, GEOMETRY, AND PACKAGING

A. Membrane Material Selection

Membranes used for separation are thin selective barriers. They may be selective on the basis of size and shape, chemical properties, or electrical charge of the materials to be separated. As discussed in previous sections, membranes that are microporous control separation predominantly by size discrimination, charge interaction, or a combination of both, while nonporous membranes rely on preferential sorption and molecular diffusion of individual species. This permeation selectivity may, in turn, originate from chemical similarity, specific complexation, and/or ionic interaction between the permeants and the membrane material, or specific recognition mechanisms such as bioaffinity.

A membrane material should meet several criteria: it should be chemically and physically stable under anticipated operating conditions, have the permselectivity required for a given process design, and be conveniently fabricated into membrane form. Polymers are the most frequently used membrane materials as they offer a wide spectrum of properties. Specialty membranes made of inorganic materials such as ceramics, metals, and carbon are also available. Their ability to withstand extreme temperatures and harsh chemical conditions enables their deployment in applications not addressed by polymeric membranes. Membranes used in the life sciences are designed to contact delicate biological or biochemical materials; a high degree of biocompatibility and hydrophilicity is necessary to minimize nonspecific interaction and the consequent degeneration in membrane performance or damage to the biological material.

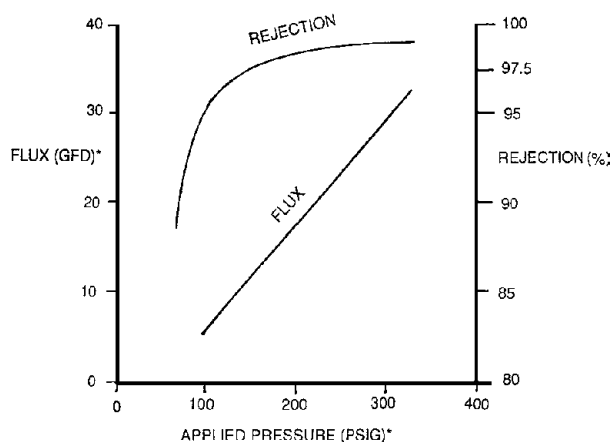


FIGURE 4 Flux in GFD (gal/ft²/day) and rejection of NaCl at 25°C for atmospheric pressure permeate with increasing applied feed pressure with a 5000 mg/L salt feed. The membrane is an asymmetric polyamide.

In certain cases, the separation medium of a membrane is a liquid that is immiscible with the feed stream. The very high permeability of liquids relative to solid materials offers a productivity advantage. Usually the selectivity of liquids derives from differential partitioning of permeants. Liquids may also be used as a solvent for specific complexing agents that do not form membranes themselves. Finally, transient deposits of colloids can be used as selective barriers in the so-called dynamic membranes, which offer very high productivities when moderate degrees of separation are adequate.

B. Membrane Structure and Geometry

For membrane separation processes, productivity is often measured in terms of permeation flux. High fluxes are achieved by using thin membranes. The invention and widespread use of several types of membranes with sub-micron separation layers is largely responsible for the phenomenal growth of applied membrane technology. In “asymmetric” membranes, the structural density changes from one surface of the membrane to the other, with the part of highest density being the functional separation layer. “Composite” membranes have a multilayered construction: a thin separation barrier supported by a relatively thick, nonselective substrate. Both types of membranes are used extensively for industrial separations of low-molecular-weight substances.

Another means of classifying membranes is according to their ability to retain substances of different sizes. Some membranes are capable of size discrimination at the molecular level—for example, with gases or liquids—while others exhibit selectivity toward particles of microscopic dimensions. As will be shown in the following sections, membrane processes in conjunction with appropriate membranes can achieve separation over a broad size spectrum.

“Homogeneous” membranes have a uniform structure (even if they are microporous or nanoporous) throughout their thickness. Membranes used as depth filters generally have this structure. They are also preferred when the application calls for membranes with a nondirectional character, as in electrodialysis (*q.v.*), when the material is difficult to fabricate into asymmetric or composite membranes, or when high fluxes are not important, as in controlled release. [Table IV](#) lists separation membranes by materials and structural features.

Membranes used in nonseparation applications may have special structural requirements. Examples are membranes that serve as flow-through chemical reactors (*q.v.*), in which reactants are converted to products by contact with catalysts inside the pores of the membrane, or as a reversible adsorption matrix based on biospecific inter-

actions (i.e. affinity separation) (*q.v.*). These applications call for a high internal surface area in the membrane, such as that afforded by a finely porous, open-cell morphology. Membrane thickness does not affect productivity directly in such cases. Indeed, thicker membranes may be preferred because they permit longer residence times for more complete reaction or capture of target species, so long as the flow of reactants and products is not unduly hindered.

Two common membrane geometries are flat-sheet and tubular (including hollow fibers). Flat-sheet membranes are made by casting, coating, or extrusion. A nonwoven fabric backing is often used to provide mechanical reinforcement. Tubular and hollow-fiber membranes are made by spinning or extrusion, depending on diameter. Inducing phase separation in a polymer solution—either thermally or by controlled mixing with a nonsolvent—typically forms the microporous structure. Liquid membranes are either microdroplets in the form of emulsions prepared and handled by liquid-liquid extraction equipment, or immobilized in a porous support to assume a stable physical form.

C. Membrane Modularization and Packaging

Synthetic membranes are delicate and fragile by nature. There are instances in which individual sheets of membrane are used in holders or housings, particularly in a laboratory setting. Careful handling and controlled environments are essential to protect the membrane from damage or contamination.

Independent of which type of membrane is being used, a large amount of membrane area must be accommodated in an efficient system. Since compactness is important, clever designs have evolved to incorporate large amounts of membrane area in efficient modules. Virtually all membranes used industrially are packaged as modules. Packaging also protects the membranes from damage, and facilitates changes in capacity by changing the number or size of devices. Secondary factors such as the need to control external phase fluid dynamics are sometimes important in practical module selection when phenomena known as concentration polarization and fouling must be dealt with. (see Section B.1b). Flat-sheet membranes may be packaged as spiral-wound elements or pleated cartridges, or used in single sheets in plate-and-frame modules. Tubular and hollow-fiber membranes are usually formed into bundles secured by potted tube sheets at one or both ends and housed in a cylindrical shell. Some common commercial module designs are shown in [Fig. 5](#).

The choice of a preferred module design is determined by technical and economic factors specific to each application. Two key variables govern cost: the productivity per unit membrane cost, and the life expectancy of the

TABLE IV Membrane Classification

Separation material	Structure	Morphology	Geometry	Methods of fabrication	Typical applications
Polymers	Homogeneous	Microporous	Flat-sheet, tubular, hollow fiber	Phase-inversion casting or spinning, sintering, track-etching, biaxial stretching, anodizing	Microfiltration, membrane distillation, affinity separation
		Nonporous	Flat-sheet, hollow fiber,	Extrusion, casting	Diaysis, electrodialysis, controlled release
	Asymmetric	Microporous	Flat-sheet, tubular, hollow fiber	Phase-inversion casting or spinning	Microfiltration, ultrafiltration, membrane reactors
		Nonporous, skinned on microporous substrate	Flat-sheet, tubular, hollow fiber	Phase-inversion casting or spinning	Reverse osmosis, gas separation, pervaporation, perstraction, membrane reactors
	Composite	Nonporous barrier on microporous substrate	Flat-sheet, hollow fiber	Direct coating, interfacial polymerization, plasma polymerization	Reverse osmosis, gas separation, perstraction
Inorganic (ceramic, metal, carbon)	Isotropic or asymmetric	Microporous	Tubular, multichannel monolithic	Sol-gel inversion, sintering, calcining, anodizing, carbonizing of polymeric precursors	Microfiltration, ultrafiltration, membrane reactors
Liquid	Continuous	Liquid immobilized in microporous substrate	Flat-sheet, hollow fiber	Impregnation	Membrane extraction, gas separation, coupled transport
	Emulsion	Micellar	Microdroplets	Single- or multistage emulsification	Emulsified liquid membrane extractions
Colloidal (dynamic)	Transient gel-like coating	Colloidal barrier layer on porous substrate	Tubular	Formed in place during operation	Ultrafiltration, reverse osmosis
Gas	Continuous	Gas trapped in microporous by external liquid	Flat sheet, hollow fiber	Formed-in-place	Recovery of volatile substances from liquids

membrane device. Cost decreases as processing capacity per module increases. This consideration favors devices with high packing density, or large membrane area per unit module volume. Another consideration is to prolong the useful life of the module, hence reducing the frequency of membrane replacement. Membrane lifetime is affected mainly by the interaction between the membrane and the feed material, and by operating conditions that control the rate of reversible and permanent performance degradation. All membrane modules have finite lifetimes and ultimately require replacement. Some of them are designed to be disposable devices intended for single use; their values are less dependent on length of service than the need to maintain a consistent level of performance.

High flow rates across the membrane surface help reduce the accumulation of solutes rejected by the membrane (referred to as concentration polarization) and impurities lodged on the membrane surface (i.e. “fouling”). (See Section 1bii.) Tubular membranes and flat-sheet membranes installed in thin channel plate-and-frame

stacks readily accept high flow rates; they are also more conveniently cleaned by mechanical means or by disassembly. However, these configurations provide relatively low membrane area per unit module volume. Spiral-wound modules have a higher packing density. Capillary-like hollow fibers are prone to fouling, but they offer the highest packing density. Hollow fibers with internal diameters from about 0.1 to 1 mm combine relatively high packing density and the flexibility of lumen- or shell-side feed at moderate flow rates.

Special considerations apply to the design of products intended for single use but to allow analysis to be conducted rapidly and on a relatively small scale. For example, a membrane may be packaged in a small holder that attaches to the tip of a syringe to filter milliliter quantities of solution. Centrifuge tubes may have a membrane partition built in so that biological samples may be separated or rinsed with buffer solution as a part of centrifugation (Fig. 6a). With the advent of biotechnology, large-scale screening procedures demanded dramatic increases

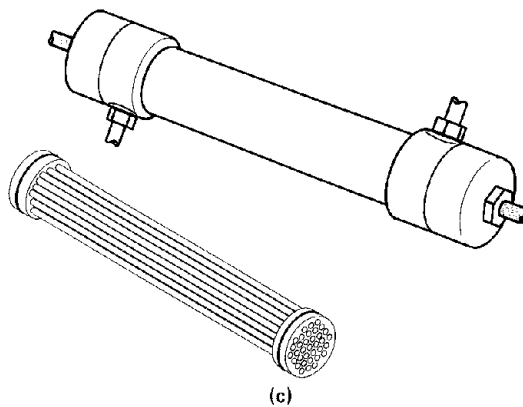
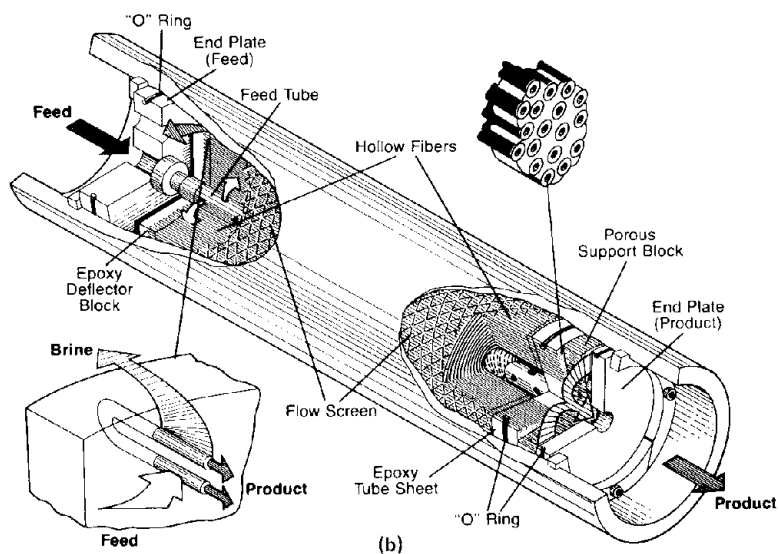
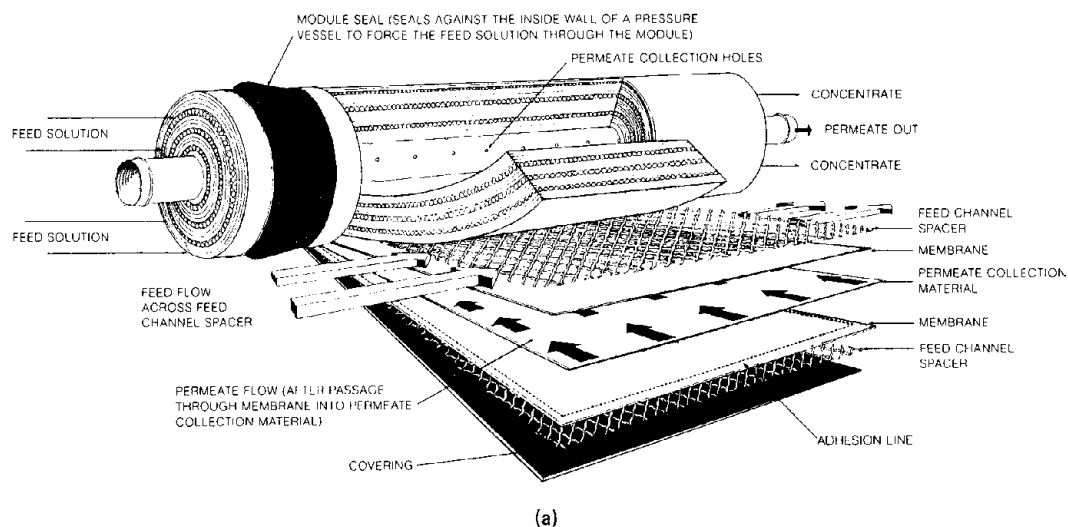


FIGURE 5 Membrane module design. (a) Spiral-wound (Koch Membrane Systems); (b) hollow-fiber (Du Pont); (c) tubular (generic); (d) plate-and-frame; (e) pleated cartridge (Millipore). [Figure 2(d) from Strathmann and Chmiel (1985)].

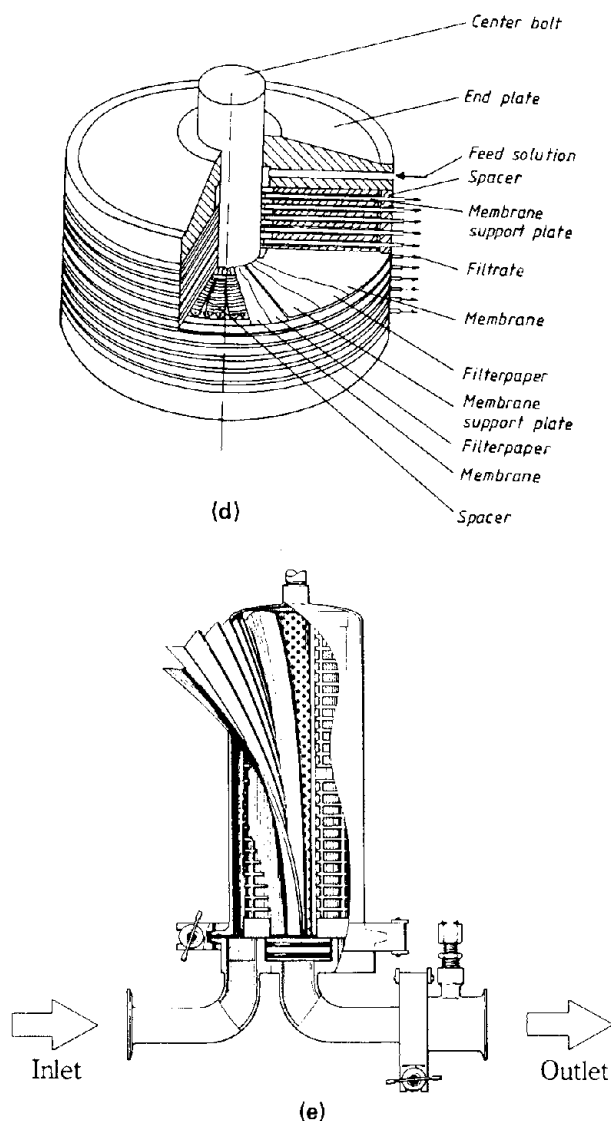


FIGURE 5 (continued)

in productivity. Membranes are packaged in the form of multiwell plates designed for automated equipment and methods of analysis (Fig. 6b).

Specialty membrane devices used as sensing elements and electrode components are often built permanently into instruments. Diagnostic or medical devices are often single-use disposable items.

III. GAS SEPARATIONS

A. Overview of Separation Processes Involving Gases and Vapors

As one considers gas and vapor feeds instead of liquids, new issues emerge. Carefully drying of micro-, ultra-, or

nanofiltration membranes preserves their basic pore size distributions. These dried membranes can be used with gaseous streams. Indeed, ambient temperature sterilization of air is possible with a membrane that removes particulates less than the size of a virus ($\sim 0.1 \mu\text{m}$). In microelectronics and pharmaceuticals, where not only microbes but also their fragments can cause problems, this is obviously an advantage. In general, however, membrane separation is applied to gas or vapor mixtures to achieve a *molecular* separation between the stream components.

An even wider diversity of mechanisms can effect molecular level separations of gases and vapors as compared to liquid mixtures. The simplest approach involves applying a transmembrane mixed gas pressure across a membrane. Depending upon the structure of the membrane, this process may or may not cause separation of the coperming components. For porous membranes, the size of the pores relative to the mean free path of the molecules under the conditions of the feed and permeate will determine the outcome. If the gas molecules collide preferentially with each other *instead* of the pore wall (i.e., the pore diameter exceeds the bulk mean free path), viscous flow applies, and no separation occurs. On the other hand, if the mean free path between collisions in a normal bulk gas phase of equal pressure exceeds the pore size of the membrane, separation occurs. This process, termed “Knudsen diffusion,” is promoted by operation at low pressures or by using membranes with small pores at elevated pressures. The more rapidly moving low molecular weight gas executes more frequent diffusional steps, since it hits the wall more frequently. The ratio of wall collisions in this limit scales with the inverse square root of penetrant molecular weight. Therefore, the Knudsen selectivity equals the inverse square root of the molecular weight ratio of the largest to smallest gas (Koros and Pinnau, 1994). This principle was used for isotope enrichment on the Manhattan Project, but it is uneconomical for commercial separation applications.

B. Practical “Contender” Membranes for Gas and Vapor Separations

Besides Knudsen diffusion, permselective transport of gases can occur by various mechanisms involving molecular scale interactions of the sorption-diffusion type. These can be broadly classified into three groups as described below and pictured in Fig. 7.

1. “Simple” Sorption-Diffusion Mechanism

The sorption-diffusion mechanism considers that some thermally agitated motions (either in the matrix or by the penetrant) provide opportunities for sorbed penetrants to

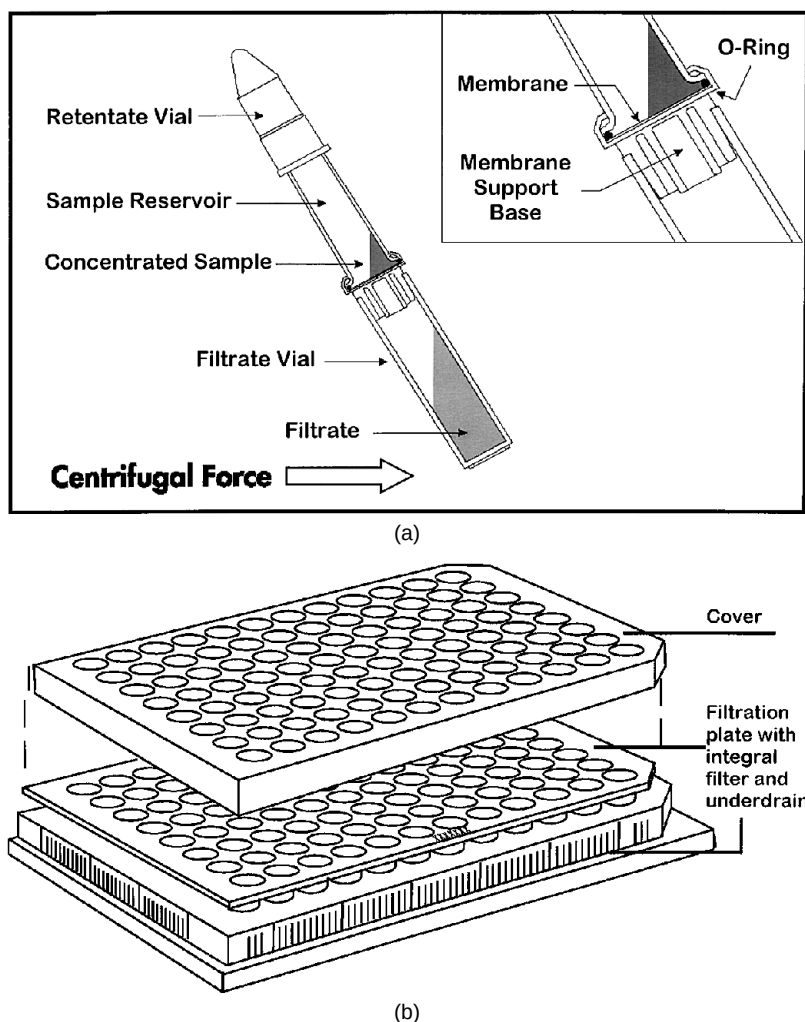


FIGURE 6 (a) Centrifuge-tube membrane filter (Millipore Corporation). (b) The 96-well plate with membrane sealed to individual cavities with integral underdrain receptacles (Millipore Corporation).

diffuse from the upstream to the downstream face of a membrane (Fig. 1). Like reverse osmosis, the driving force for gas separation is a chemical potential difference related to the concentration difference imposed between the feed and permeate sides of the membrane. For gas separation, this chemical potential difference arises from a partial pressure (or fugacity) difference of the permeating species between the upstream and downstream membrane faces (Koros and Hellums, 1989). Such membranes can be further sorted into three groups: polymeric solution-diffusion, molecular sieving, and selective surface flow.

In any case, the “permeability,” P_A , of a given gas (A) in a membrane material simply equals the pressure-and-thickness-normalized flux. This parameter provides the overall measure of the ease of transporting the gas through the material.

$$P_A = [\text{flux of A}][L]/[\Delta p_A]. \quad (11)$$

In terms of Eq. (1), the driving force is Δp_A and the resistance, $\Omega_A = L/P_A$. Although the effective skin thickness L is often not known, the so-called permeance, P_A/L can be determined by simply measuring the pressure normalized flux, viz., $P_A/L = [\text{flux of A}]/\Delta p_A$, so this resistance is known. Since the permeability normalizes the effect of the thickness of the membrane, it is a fundamental property of the polymeric material. Fundamental comparisons of material properties should be done on the basis of permeability, rather than permeance. Since permeation involves a coupling of sorption and diffusion steps, the permeability is a product of a thermodynamic factor, S_A , called the solubility coefficient, and a kinetic parameter, D_A , called the diffusion coefficient.

$$P_A = [S_A][D_A]. \quad (12)$$

The coefficients in Eq. (12) are themselves complex functions that depend upon the type and amount of other sorbed

Type	Typical Features of Current Primary Membrane Types
A. Asymmetric polymeric solution-diffusion	Thin selective skin layer (0.1 μm) Highly porous support 250 μm OD $D_i = f_i \lambda_i^2 / 6$ diffusion step
B. Molecular sieving (zeolite or carbon)	Reverse selective skin layer (0.1-10 μm) Highly porous support layer (wide range of sizes & morphologies)
C. "Reverse selective" surface diffusion	Reverse selective skin layer (1-5 μm) Highly porous ceramic or carbon support 7600 μm OD

Type	Typical Features of Current Primary Membrane Types
D. Complexing & Reactive	e.g. O_2 carrier facilitated membranes N_2 O_2 Air e.g. palladium alloy membranes for H_2 H_2 $\text{H}_2 \rightarrow 2\text{H}$
E. Proton exchange (PEM) (e.g. Nafion®)	750 μm 0.5 $\text{O}_2 \rightarrow \text{H}_2\text{O}$ PEM H^+ $\text{H}_2 \rightarrow$ Load Fuel Cell
F. Solid Oxides	1000 μm CH_4 Air O^{2-} 2$\text{H}_2\text{O} + \text{CO}_2$ Load Fuel Cell CH_4 Air O^{2-} 2e^- $\text{C}_2\text{H}_4 + \text{H}_2\text{O}$ Oxidative Coupling Membrane Reactor

FIGURE 7 Practical gas separation membrane types.

penetrants near the permeating penetrant. Temperature is also an important factor which activates the diffusion jumps and moderates the thermodynamic interactions between the sorbed penetrants and the matrix.

The separation factor for component A vs B, α_{AB} , is defined in terms of the downstream and upstream mole fractions (Y) of components A and B:

$$\alpha_{AB} = [Y_{A1}/Y_{B1}]/[Y_{A2}/Y_{B2}]. \quad (13)$$

Under ideal conditions with a negligible downstream pressure of both components, the separation factor can be equated to the *ideal membrane selectivity* factored into its mobility and solubility controlled contributions, viz.,

$$\alpha_{AB}^* = P_A/P_B = \left[\frac{D_A}{D_B} \right] \left[\frac{S_A}{S_B} \right]. \quad (14)$$

mobility solubility
controlled controlled
factor factor

For a defect-free ideal membrane, the selectivity is independent of thickness, and either permeability ratios or permeance ratios can be used for comparison of selectivities of different materials. Nonideal module flow patterns, defective separating layers, impurities in feeds, and other factors can lower the actual selectivity of a membrane compared to tabulated values based on ideal conditions (Koros and Pinnau, 1994).

Currently, all commercial gas and vapor separation membrane are either glassy or rubbery polymers (Spillman, 1989; Puri, 1996; Meindersma and Kuczyński, 1996). Glassy materials generally derive permselectivity from their ability to separate gases based on subtle differences in penetrant size with minor contributions from the solubility controlled term. Rubbery materials, on the other hand, generally derive permselectivity from favorable solubility selectivity with minor contributions from the mobility term. In both cases, transport is postulated to occur upon the creation, next to the penetrant molecule, of a transient gap of sufficient size to accommodate the penetrant, thereby permitting a diffusion step (Fig. 7A) (Koros and Hellums, 1989). These transient gaps form and fade throughout the polymer due to thermally induced motions of the polymer chain segments. Polymeric membranes tend to be more economical than other materials and thus dominate traditional gas separations. The low cost of polymeric membranes results from their ability to be easily formed into hollow asymmetric fibers or spiral wound modules, due to their segmental flexibility and solution processability. Extremely thin (less than 0.1 μ) separating layers (Fig. 2) are currently achievable with such materials (Zolandz and Fleming, 1992). The segmental flexibility of polymeric membranes that makes them economical to prepare, in fact, limits their discriminating ability for similarly sized

penetrants (Singh and Koros, 1996). Moreover, the loss in performance stability at high temperature, at high pressure, and in the presence of highly sorbing components limits the wider scale use of these otherwise versatile membranes.

The values of permeability coefficients for He, O₂, N₂, CO₂, and CH₄ in a variety of “dense” (isotropic) polymer membranes and the overall selectivities (ideal separation factors) of these membranes to the gas pairs He/N₂, O₂/N₂, and CO₂/CH₄ at 35°C have been tabulated in numerous reviews (Koros and Hellums, 1989; Koros, Fleming, and Jordan *et al.*, 1988; Koros, Coleman, and Walker, 1992). Moreover, several useful predictive methods exist to allow estimation of gas permeation through polymers, based on their structural repeat units. The values of the permeability coefficients for a given gas in different polymers can vary by several orders of magnitude, depending on the nature of the gas. The values of the overall selectivities vary by much less. Particularly noteworthy is the fact that the selectivity decreases with increasing permeability. This is the well-known “inverse” selectivity/permeability relationship of polymer membranes, which complicates the development of effective membranes for gas separations.

Typically, membranes with high gas permeabilities and a low selectivities are comprised of “rubbery” polymers, i.e., $T_g < T$, where T_g is the glass-transition temperature of the polymer and T is the temperature at which the permeability is measured. Rubbery polymers are characterized by high intrasegmental mobility, whereas glassy polymers exhibit the opposite characteristics. An interesting exception to this rule is poly[1-(trimethylsilyl)-1-propyne] (PTMSP), which is a rigid glassy polymer but nevertheless exhibits the highest intrinsic gas permeability of all known synthetic polymers. The high permeability of PTMSP has been found to be due to an exceptionally large free volume that appears to provide a system of interconnected microporous domains of about 5–15 Å in size within the PTMSP matrix (Stern and Koros, 2000). This material, therefore, appears to border on nanoporosity in its properties.

The above-mentioned “inverse” selectivity/permeability relationship of polymers has been summarized by Robeson by means of log–log plots of the overall selectivity versus the permeability coefficient, where A is considered to be the more rapidly permeating gas. These plots were made for a variety of binary gas mixtures from the list He, H₂, O₂, N₂, CO₂, and CH₄, and for a large number of rubbery and glassy polymer membranes. Such representations, shown in Fig. 8 and Fig. 9 are often referred to as “upper bound” plots (Robeson, 1991). The “upper bound” lines clearly show the “inverse” selectivity/permeability relationship of polymer membranes. While these plots were prepared in 1991, only small advances have been made to push the upper bound higher since that time.

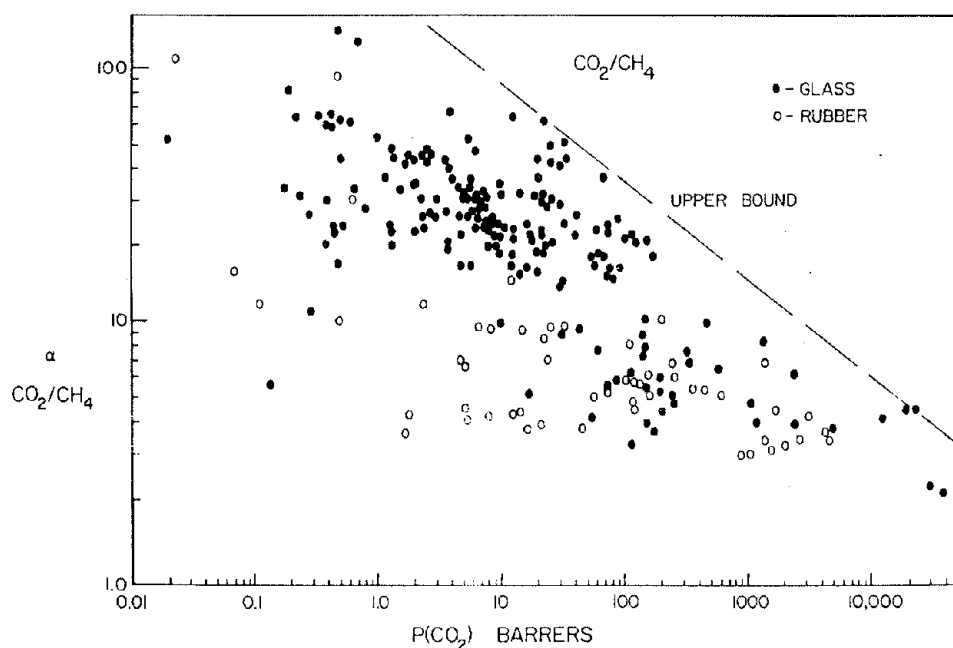


FIGURE 8 Literature data for CO_2/CH_4 separation factor vs CO_2 permeability.

The factors affecting the selectivity and permeability of polymer membranes to different gases are best discussed on the basis of Eqs. (12) and (14). As noted in Eq. (12), the permeability coefficient, P , of a penetrant gas in a polymer membrane is the product of a (concentration-averaged) diffusion coefficient, D , and of a solubility coefficient,

S . The diffusion coefficients of gases in glassy polymer membranes are strong functions of the penetrant gas concentration in the membranes (or of the gas pressure), and depend also on polymer morphology (crystallinity, orientation), crosslinking, and chain mobility. The chain mobility depends, in turn, on the polymer free volume, the

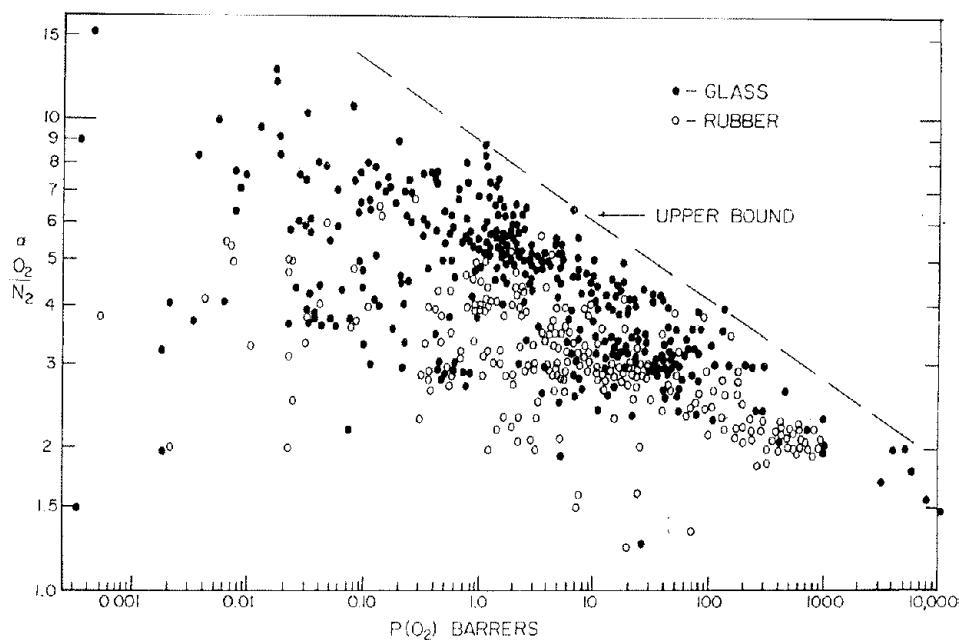


FIGURE 9 Literature data for O_2/N_2 separation factor vs O_2 permeability.

free-volume distribution, and the temperature (Koros and Hellums, 1989; Prasad, Notaro, and Thompson, 1994; Kesting and Fritzche, 1993; Gas Processors, 3132-84). The diffusion coefficients of the components of a gas mixture may also depend on composition. The solubility coefficients depend primarily on unrelaxed volume in the polymer, the penetrant condensability, and to a lesser degree upon penetrant-polymer interactions (Spillman, 1989; Zolanz and Fleming, 1992; Koros and Hellums, 1989). Therefore, the permeability and selectivity coefficients depend on all of the above factors in view of Eq. (14), but the overall selectivity of glassy polymer membranes depends mainly on the diffusivity selectivity. This diffusivity selectivity can vary by an order of magnitude or more depending on the nature of the membrane and of the gas pair under consideration. The diffusivity selectivity, and hence the “sieving” ability, of glassy polymer membranes is significant even when the difference in the sizes of penetrant molecules is very small. For example, the “kinetic” diameters of O₂/N₂ pair differ by only 0.18 Å (3.46 vs 3.64 Å; Koros and Hellums, 1989) while the He/CH₄ pair shows a “large” difference of 1.2 Å in kinetic diameters.

Fractional free volume, comprised of the average unoccupied space within the polymer matrix, are the most commonly used parameters for correlating permeabilities, and as noted earlier, group contribution methods exist to assist in such estimations (Park and Paul, 1997; Robeson, Smith, and Langsam, 1997). Unlike rigid glassy polymers, *rubbery* polymers have a low ability to discriminate between penetrant molecules of different sizes and shapes, due to the high segmental mobility of such polymers. As a result, the overall selectivity of such membranes to different gases is controlled mainly by the *solubility selectivity*. The solubility of gases in polymers commonly increases with increasing critical temperature, T_c , of the penetrant gases—hence, the solubility selectivity of rubbery polymer membranes to a gas pair will be larger the greater the difference in the T_c of the two gases. Therefore, rubbery polymer membranes are well suited for the separation of easily condensable organic vapors with high T_c 's from light gases with low T_c 's, such as the components of air.

The solubility selectivity of a membrane for a specific gas pair could be increased (in principle) by inducing specific interactions between the polymer and the more soluble component of the gas pair. For example, the substitution of certain polar groups in some rubbery polymers has been found to increase their solubility selectivity for CO₂ relative to CH₄ (Story and Koros, 1991; Koros, 1985). Unfortunately, the increase in the polarity of a polymer also tends to increase its chain packing density, and as a result, decreases the gas diffusivity in membranes made from that polymer.

The above discussion raises a question regarding the degree to which the selectivity of polymer membranes to specific gas pairs can be enhanced by structural modifications without significant loss in permeability. The question posed is whether the lines in selectivity/permeability plots, such as the dashed lines in Figs. 8 and 9, have an upper limit. The consensus of these analyses (Singh and Koros, 1996; Park and Paul, 1997; Robeson, Smith, and Langsam, 1997; Alentiev, Loza, and Yampol'skii, 2000; Freeman, 1999) generally support the preceding qualitative conclusions noted above that such an upper bound does exist for each gas pair using polymers that can be processed by conventional solution casting methods. Specifically, it appears that the segmental flexibility of polymeric membranes that makes them economical to prepare, in fact, limits their size and shape discriminating ability for similarly sized penetrants.

Molecular sieving materials are an alternative to polymers. Like glassy polymers, such media rely primarily on differences in molecular size to achieve separation, but the detailed diffusion step is rather different in the two cases. Molecular sieve membranes are ultramicroporous, with sufficiently small pores to exclude some penetrants while allowing others to pass through (Fig. 7B). These rigid membranes show extremely attractive permeation performance (Morooka and Kusakabe, 1999; Tsapatis and Gavalas, 1999) and maintain stability when exposed to adverse conditions (high temperature, pressure, highly sorbing components) that can cause polymeric membranes to plasticize. Under ideal conditions, minimum effective thickness layers similar to those achievable with polymeric membranes (~ 0.05 – $0.2 \mu\text{m}$) can be obtained with some molecular sieving materials. Unfortunately, such membranes are difficult to process, are fragile, and expensive to fabricate into modules; thus, they are not commercially significant today except in niche applications.

As noted above, glassy polymers and molecular sieving materials preferentially permeate the smallest component in a mixture compared to larger sized components in the mixture. In certain separations, it may be advantageous to permeate the larger sized penetrant and retain the smaller component. These separations can be potentially achieved using “surface selective flow” membranes (Rao and Sirkar, 1993, 1997). While rubbery polymers show this property, the selectivity achievable is generally not impressive except when comparing a highly condensable component and a supercritical gas like air. On the other hand, uniformly nanoporous membranes have been reported that show a high degree of such “reverse selectivity.” These nanoporous materials work by the selective adsorption of the more strongly adsorbing components on to the pore surface followed by surface diffusion of the adsorbed molecules across the pore (Fig. 7C). The

adsorbed molecules create a hindrance to transport of smaller nonadsorbed species through the void space in the pores. These membranes have reasonable transport properties and can be attractive if the desired separation cannot be achieved by conventional methods. Pilot-scale membrane modules using surface selective flow for hydrogen enrichment have recently been tested (Anand, Langsam, Rao, and Sircar, 1997).

2. “Complex” Sorption-Diffusion Membranes

These membranes are similar to the “simple” sorption-diffusion membranes, but involve some additional phenomena as well as simple penetrant dissolution and diffusion. Two types can be identified: (i) facilitated transport for various gas types, and (ii) palladium and related alloys for hydrogen.

Facilitated transport membranes involve a reversible complexation reaction in addition to simple penetrant dissolution and diffusion. The penetrant sorbs into the membrane and diffuses down the conventional concentration gradient, or it can react with complexation agent or carrier agent and diffuse down a concentration gradient of a carrier–gas complex (Fig. 7D). The later transport mechanism is not accessible to other penetrants that do not react with complexation agent. Transmembrane chemical potential difference, is of course, still the driving force for permeation. These membranes are highly selective and can potentially achieve high permeabilities at low concentration driving force (Way and Noble, 1992; Cussler, 1994). These membranes are configured either as an immobilized liquid film, a solvent swollen polymer, or a solid polymer film containing reactive functional groups. The main disadvantage of these membranes is the potential lack of stability: the membranes can dry out or the carrier species can be lost. Until the issues relating to stability are resolved, facilitated transport membranes are unlikely to be used for large-scale gas separations. Besides gas separations such carrier facilitated membrane can be used in liquid separations or ion fractionation, but similar instabilities have plagued these cases as well until recently (Ho, 2000).

Palladium-based membranes are highly selective to hydrogen (Ma, 1999; Wood, 1968) that can also be interpreted in terms of a “complex sorption-diffusion” mechanism. In this case, permeation of hydrogen through Pd membranes involves the dissociative adsorption of hydrogen onto the surface. A palladium hydride is believed to form with partial covalent bonds (something between true chemical binding and interstitial alloys) (Glasstone, 1950). This initial step is followed by the transition of atomic hydrogen from the surface into the bulk of the metal, followed by atomic diffusion through the mem-

brane. The above mentioned steps then occur in reverse order at the downstream membrane face (Fig. 7D). Since the permeation process is controlled by the diffusion of atomic hydrogen, the flux is proportional to the difference of the square root of pressures of hydrogen (Sievert’s law). Palladium alloys are often preferred, because pure palladium tends to become brittle after repeated cycles of hydrogen adsorption and desorption. These membranes are typically used as membrane reactors, which combine some reaction leading to generation of hydrogen along with hydrogen separation in a single unit. For certain chemical reactions, e.g., propane dehydrogenation, natural gas steam reforming, these membrane reactors show good transport properties as well as temperature resistance (Ma, 1999). However, there are still considerable difficulties in preparing these membranes for economic operation on a large scale.

3. Ion-Conducting Membranes

Organic polymeric and ceramic ion conducting materials can be used in formulating membranes for some specialty gas separation application. The most important of these are solid oxides and proton exchange types (Fig. 7E and 7F).

The solid oxide materials are permeable to oxygen ions and can be further divided into two classes: mixed ionic electronic conductors and purely oxygen ion conductors. The mixed ionic electronic conductors are capable of conducting both oxygen ions and electrons. These mixed ion-conducting materials are being studied being for processes where oxygen or oxygen ions are required. The oxygen permeation process through oxygen ionic conducting membranes involves three mass transfer steps: electrochemical surface reactions at the two gas-membrane interfaces and oxygen ion transport through the bulk oxide. These materials are mostly oxides called perovskite and have the generic formula ABO_3 , where A is a large cation with a 12-fold coordination and B is a smaller cation with a sixfold coordination with oxygen ions. When the ions take a mixed-valence state, the partial substitution of the A site by other metal cations with lower valences can usually cause the formation of oxygen vacancies and a change in the valence state of the B ions in order to maintain charge neutrality (Ma, 1999). Oxygen ions (created by electrochemical reduction reaction on the surface) migrate via oxygen vacancies in the bulk of the membranes and then form molecular oxygen at the downstream interface by a surface oxidation reaction. These membranes have exceptionally high selectivity and high fluxes compared to polymeric membranes, and typically operate at high temperature (700°C). Despite their expected high cost, these so-called mixed ionic electronic conductors (MIEC) (Nigara, Mizusaki, and Ishigame, 1995; Balachandran,

Kleefisch, Kobylinski *et al.*, 1996; Balachandran, Dusek, Maiya *et al.*, 1997) are being considered for nonelectrochemical processes such as the production of synthesis gas from methane. In this case, as oxygen ions emerge from the downstream side of the membrane and react with methane to form syngas, the electrons that are released can diffuse back through the membrane to maintain electrical neutrality. In addition, there is work to pursue methane oxidative coupling to produce ethylene and propylene directly from methane. Other problems that need to be resolved include difficulties in proper sealing of the membranes as well as high sensitivity of membranes to the temperature gradients that can result in membrane cracking (Bessarabov, 1999). Nevertheless, these are interesting and exciting additions to the membrane spectrum.

Unlike the mixed ion conductors, solid oxides that can only conduct oxygen ions and not electrons have applications involving electrons flow through an external circuit to produce power in fuel cells (Fig. 7F). Fuel cells are electrochemical devices that directly convert available chemical free energy in a fuel by oxidizing the fuel, typically hydrogen, methanol, or some other hydrocarbon into electrical energy. One type of fuel cell uses oxygen-conducting materials (Lin, Wang, and Han, 1994). Here oxygen ionizes to form oxygen ions and the oxygen ions diffuse through the membrane to react with a hydrocarbon on the other side to form CO_2 and H_2O . As a result, electrons flow back through the external circuit to maintain electrical neutrality, thus providing electrical power. To provide adequate oxygen fluxes, high temperatures are required ($>650^\circ\text{C}$).

A second type of fuel cell is based on the proton-exchange membranes described below (Heitner-Wirguin, 1996). Unlike the solid oxide membranes, proton exchange membranes offer the opportunity to operate at lower temperatures than the solid oxides. Proton exchange membranes (Fig. 7E) are the mirror image of the oxygen ion conducting solid oxide membranes described earlier (not the MIEC), since they only conduct protons and not electrons. These can be polymeric or inorganic, and the most popular of these is Nafion, a perfluorinated sulfonic acid polymer. Other sulfonic acid containing materials are also under study. Addition of water to these sulfonated polymers causes the hydrogen ions on the SO_3H groups to become mobile. It is proposed that proton conductivity in these materials is a result of two different mechanisms (Pivovar, Wang, and Cussler, 1999). In one mechanism the protons add on to one side of a water molecule and hop off the other side to a different water molecule, and so on. The other mechanism is somewhat like the facilitated transport mechanism described earlier. Specifically, the proton combines with a solvent molecule to yield a complex and then the complex diffuses through the membrane.

For fuel cells, the assembly consists of an ion-conducting film sandwiched between two platinum based electrodes. Hydrogen fuel is typically supplied to the anode, while the oxidant is supplied to the cathode. Hydrogen is dissociated at the anode, catalyzed by the platinum, to yield electrons and hydrogen ions. The hydrogen ions migrate through the proton exchange membrane while electrons travel to the cathode through an external circuit. The protons and electrons react with oxygen at the cathode to produce water and heat. The driving force for the reaction manifests itself in the voltage that drives the electrons through the external circuit (Singh, 1999).

The biggest advantages of fuel cells over conventional automotive energy production is the efficiency (twice as high internal combustion engines) and near zero emissions. There are, however, still a number of technical hurdles that need to be overcome before this process is commercialized; these hurdles include how the fuel may safely be supplied and how the cost of the catalyst can be minimized.

C. Strategies to Deal with Gas Separation Membranes Shortcomings

While concentration polarization and fouling are the main challenges facing membranes for liquid separations, gas separation systems are limited more generally by lack of durability and adequate selectivity. Therefore, a generic technical challenge typical of most potential applications of gas separation membranes includes finding ways to achieve higher permselectivity with at least equivalent productivity. Maintaining these properties in the presence of complex and aggressive feeds is the second challenge that must be balanced against cost in all cases. The relative importance of each of these requirements varies with the application. Of these requirements, selectivity (or separation efficiency) and permeation rate (or productivity) are clearly the most basic. The higher the selectivity, the more efficient the process, the lower the driving force (pressure ratio) required to achieve a given separation, and therefore the lower the operating cost of the membrane system. The higher the flux, the smaller the required membrane area and, therefore the lower the capital cost of the membrane system.

The preceding discussion of gas separation membrane types illustrates the large number of options available. A correspondingly large number of potential opportunities for gas separation membranes exist, but economics ultimately must dictate which membrane approach, if any, should be used in each application. Moreover, the key requirements of durability, productivity, and separation efficiency must be balanced against cost in all cases. The current spectrum of applications of gas separation membranes

include nitrogen enrichment, oxygen enrichment, hydrogen recovery, acid gas (CO_2 , H_2S) removal from natural gas, and dehydration of air and natural gas. In addition, fuel cells, hydrocarbon separations such as olefin–paraffin and aromatic–nonaromatic separations represent high potential new applications. All of these would benefit from more advanced membranes, or better technology to implement the membrane types mentioned in Fig. 7.

As is often the case, modifications or hybridizations of existing materials and approaches may ultimately provide the best avenue to advance the state of the art beyond the approaches discussed above. In order to understand the most attractive approaches to overcome the primary barriers to a larger range of application, it is useful to examine the current process used to form commercial hollow fiber membranes (Fig. 10) (Koros and Mahajan, 2000; Koros and Pinnau, 1994). The current membrane formation process has already been optimized to efficiently produce inexpensive membranes able to compete with alternative technologies. Therefore, deviating significantly from this process would be costly, and requires a significant justification. Fortunately, the process is quite flexible and offers considerable room for innovative adaptation. The process involves extrusion of a nascent hollow fiber of polymer solution, evaporation to produce a selective skin layer (see Fig. 10) followed by quenching, drying, and module makeup.

Overcoming the current limitation faced by gas separation membranes may be accommodated by introducing two classes of materials that lie between conventional polymers and the high-performance molecular sieving materials. These two classes, illustrated in Fig. 11 and Fig. 12, respectively, are (i) crosslinked polymers and (ii) blends of molecular sieving domains in polymers, usually referred to as “mixed matrix” materials. Such materials

may offer the vehicles for capturing new high volume opportunities mentioned above that require higher selectivities, and the ability to maintain performance in demanding environments. The first option would probably be exercised by incorporating crosslinkable groups in the polymer backbone that could be simply crosslinked in an additional step, perhaps in the fluid exchange and drying segment of the process. The second option would involve reformulating the outer skin region as is discussed below.

1. Crosslinking Approach

Crosslinking of polymer structures can overcome one of the main challenges mentioned earlier—namely maintaining membrane properties in the presence of aggressive feeds. This stabilization would be a significant advantage in high volume processing of natural gas where loss of selectivity translates to loss of valuable hydrocarbons from the nonpermeate product stream. The crosslinked structure resists swelling in the presence of plasticizing agents like CO_2 , and also promotes chemical and thermal stability (Staudt-Bickel and Koros, 1999; Rezac and Schoberl, 1999). Using the monomers shown, a crosslinkable polyimide can be formed. By using appropriate starting materials with ability to be subsequently crosslinked, the material can then be spun into hollow asymmetric hollow fibers using the scheme outlined in Fig. 11. In principle, such a material could be crosslinked in a post-treatment step by ethylene glycol using the reaction scheme outlined in Fig. 11. Recent data on crosslinked flat films formed by the above-mentioned scheme indicate that the crosslinked films maintain attractive transport properties at elevated CO_2 pressures where conventional materials typically plasticize and lose selectivity. The approach has,

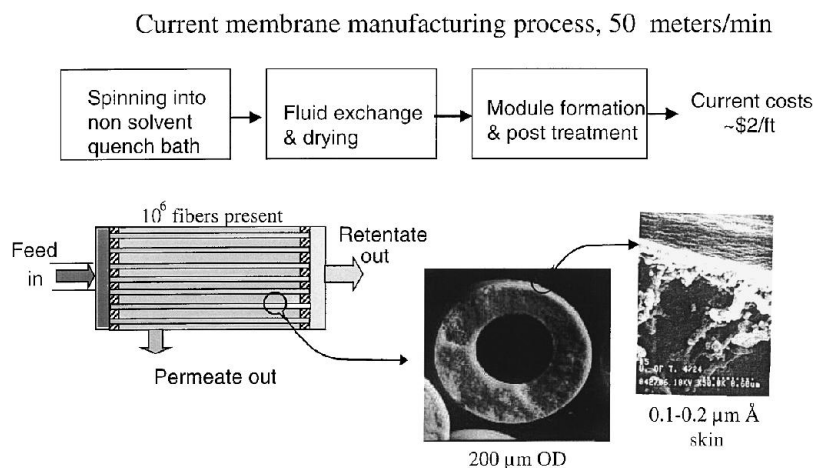


FIGURE 10 Current asymmetric hollow-fiber formation process for gas separation membranes.

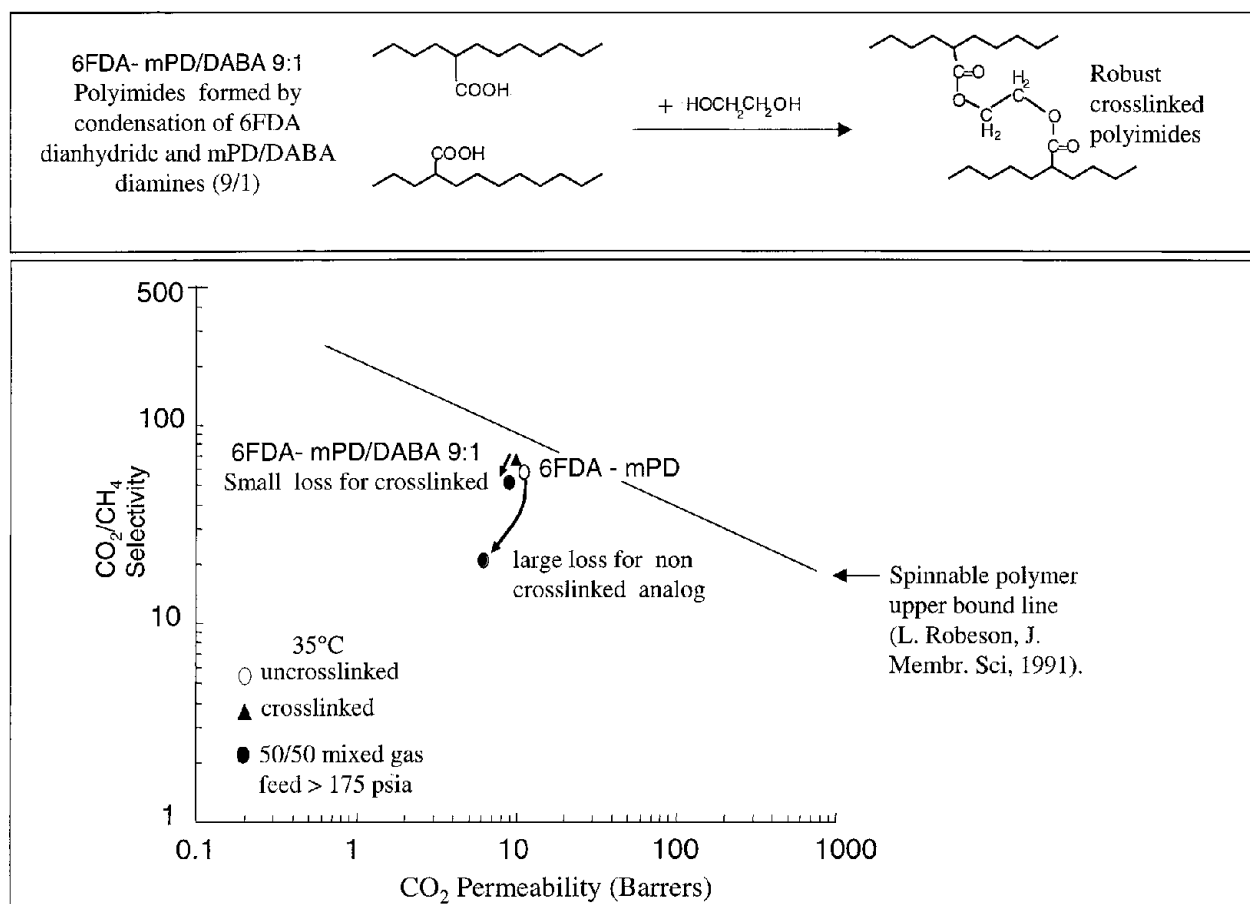


FIGURE 11 Possible strategy for stabilizing standard polymers by crosslinking.

however, not been demonstrated for actual asymmetric membranes.

2. Mixed Matrix Approach

In some cases, simply maintaining the achievable selectivity available with the current generation of gas separation polymers is not adequate. The O₂/N₂ system is an ideal example such a case, where higher selectivity membranes could reduce energy costs by as much as 20–30%. Since the raw material, air, is essentially free, this would represent a significant step forward. In this case, the so-called mixed matrix materials (Fig. 12) are attractive. Mixed matrix materials comprising molecular sieve entities embedded in a polymer matrix offer the potential to combine the processability of polymers with the superior gas separation properties of rigid molecular sieving materials. Current asymmetric composite hollow fibers consist of an inexpensive porous polymeric support coated with a thin, higher performance polymer. Similar in construction, mixed matrix composite (MMC) membranes could replace the thin, higher performance polymeric layer with

tightly packed (>20 vol %) submicron molecular sieving media, such as zeolite or carbon molecular sieves (CMS) supported within an appropriate polymeric matrix (Fig. 12). This could potentially be accomplished within much of the same cost infrastructure as used for lower performance conventional polymer. This approach has the potential to provide separation properties approaching those of high performance pure molecular sieve materials at a fraction of the cost (Mahajan, Zimmerman, and Koros, 1999). Figure 12 shows some recent data using dense films incorporating a suitable molecular sieve in a polymeric matrix and the subsequent improvement in transport properties with increasing molecular sieve loading for the oxygen/nitrogen separation (Mahajan and Koros, 1999).

Clearly, combination of the crosslinking and mixed matrix approaches to produce a robust sheath layer with embedded molecular sieve domains is a hybrid option with potential application for high selectivity needed in aggressive environments. Such materials may be the ultimate low-cost option for many of the large-scale undeveloped markets.

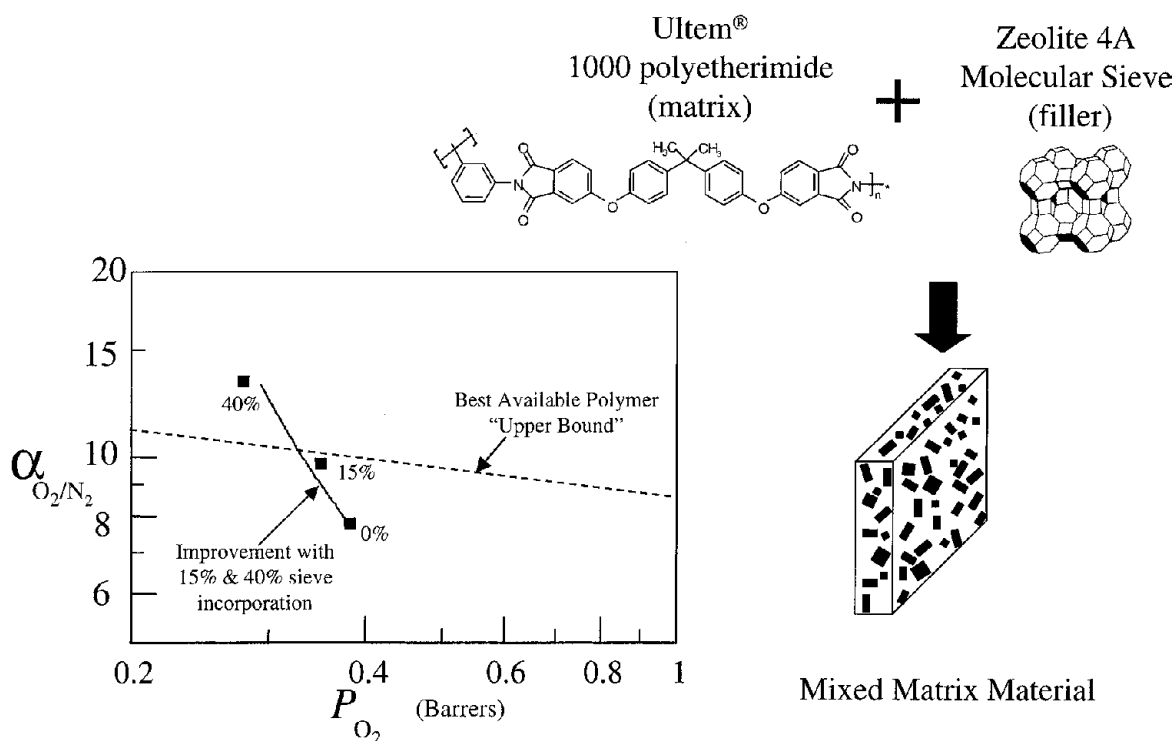


FIGURE 12 Illustration of mixed matrix strategy to exceed best available polymer performance.

While such strategies are extremely attractive, significant hurdles remain to be overcome in all cases. The crosslinking scheme needs to be tested on hollow fibers, since all reported literature is on flat-sheet membranes. Development of alternative crosslinking mechanisms is also required, as this will provide greater flexibility in implementation of this scheme. The mixed matrix work needs to be extended to polymers that are currently useful for gas separation. These materials are rigid and have issues with poor adhesion between the polymeric phase and the molecular sieving phase (Mahajan, Zimmerman, and Koros, 1999). The extension of composite spinning to spinning with sieve materials is another significant challenge to the implementation of this scheme. The polymeric materials used to mimic molecular sieves are currently processed at temperatures that would make large-scale commercialization less attractive. The development of chemistries where these materials can be produced at lower temperatures is, therefore, highly desirable.

D. Applications

The major membrane-based gas separation applications are shown in Table V. The diverse needs of these separations call for a somewhat wider range of membrane properties and module designs than is the case with liquid separations. To reflect this market and technical seg-

mentation, each major application is discussed separately below.

1. Hydrogen Separations

The first large-scale applications of membranes for gas separation were for hydrogen recovery. Hydrogen is important both as an energy resource and as a chemical feedstock. Its major uses include the synthesis of ammonia and methanol, hydrogenation of oils and fats, as reducing atmospheres in ovens, and potentially as a nonpolluting fuel. Hydrogen is produced by steam reforming of natural gas, petroleum hydrocarbons, or by electrolysis. As oil reserves become "heavier," or lower in hydrogen-to-carbon ratio with continued depletion of reserves, the overall hydrogen balance in refineries and petrochemical complexes gradually becomes increasingly deficient. Recycling hydrogen from purge streams helps reduce the load of catalytic reformers and hydrogen plants; it also minimizes supplemental purchases of hydrogen to maintain an acceptable hydrogen-to-carbon balance in petroleum refining. Some applications in the petroleum refining industry are shown in Fig. 13.

In the chemical process industry, an important application of hydrogen recovery is in ammonia synthesis purge streams. Ammonia is produced by combining hydrogen and nitrogen at high pressure and temperature in

TABLE V Gas Separation Applications

Category	Gas components	Applications	Status	Technical issues
Hydrogen	H ₂ /N ₂	Ammonia purge gas	Successful	Condensables must be removed
	H ₂ /CH ₄	Refinery hydrogen recovery	Successful	Condensables must be removed
	H ₂ /CO	Synthesis gas ratio adjustment	Successful	Condensables must be removed
	H ₂ /O ₂	Fuel Cells		
Air	O ₂ /N ₂	Nitrogen-enriched air as inerting atmosphere	Practical to 99.5%	Need more selective membranes to reach higher nitrogen purity
		Oxygen-enriched air for combustion enhancement	Various degrees of enrichment up to 50% O ₂	More selective membranes improves economics
		Home medical oxygen enrichment for respiration therapy	Successful, but small market	None
Acid gases	CO ₂ /CH ₄	Enhanced oil recovery; recover CO ₂ for reinjection	Successful	Must remove condensable hydrocarbons
		Natural gas and landfill gas sweetening	Successful	More robust and higher selectivity membranes are needed
	H ₂ S/CH ₄	Sour gas sweetening	Feasible, but no known installation	
Drying	CO ₂ /N ₂	Digester gas treatment	Successful	
	H ₂ O/HC	Hydrocarbon drying	Feasible	Hydrocarbon loss should be minimized
	H ₂ O/air	Air drying	Practical to about -10°C dew point	
Hydrocarbons	HC/air or HC/N ₂	Pollution control, volatile solvent recovery	Successful for several HCs	Permeate tends to be oxygen enriched for air case (hazard)
	HC/N ₂	Upgrading of low-BTU gas	Not yet viable	Insufficient selectivity; loss of HC into permeate
	HC/HC	Dew pointing of natural gas	Being tested	Reverse selectivity can be lost due to plugging
Helium	He/HC	Helium recovery from gas wells	Small market	Low He concentration; requires staging
	He/N ₂	Helium recovery from diving air mixtures	Feasible; small market	

the process shown in Fig. 14. The feed stream supplies the reactants and also purges trace inerts from the reactor recycle stream. Through a series of membrane units, hydrogen is recovered from the purge stream (2) and returned to the feed gas compression circuit through streams (3) and (4). The less valuable hydrogen-lean reject stream (5) is sent to the reformer as fuel. A composite membrane developed by the Monsanto Company was first used in this system. This membrane has a unique structure: an asymmetric polysulfone membrane coated with a thin layer of silicone rubber polymer. Polysulfone is selectively permeable to hydrogen, whereas the silicone rubber layer blocks the leakage of feed gas through surface pores in the polysulfone membrane to limit loss of selectivity. The efficiency and economic advantage of this process is so compelling that over the past 20 years more than 200 systems of this type have been installed.

The primary feedstock for methanol production is synthesis gas, a mixture of H₂, CO, and CO₂ from the reformer. To optimize the stoichiometry for this reaction,

the ratio between H₂/CO can be adjusted by recovering the hydrogen with a membrane system. This application is illustrated in Fig. 15. The membrane unit receives a mixture of hydrogen and methane from the purge recycle loop, and separates the hydrogen for recompression to the reactor. Removing the nonreacting methane from the recycle loop reduces circulation pumping costs and increases the concentration of the reactant gases; the result is a higher methanol yield. The methane-rich stream from the membrane unit is again used as fuel. Studies showed that the cost of the membrane system could be less than half the cost of a competitive pressure swing adsorption (PSA) system.

A similar application is the processing of fuel gas, whose major components are hydrogen (about 80%) and methane (about 20%). Asymmetric cellulose acetate membranes have been used successfully to extract the more valuable hydrogen at high purity. New membrane materials more resistant to harsh conditions will accelerate the application of other H₂ recovery schemes for

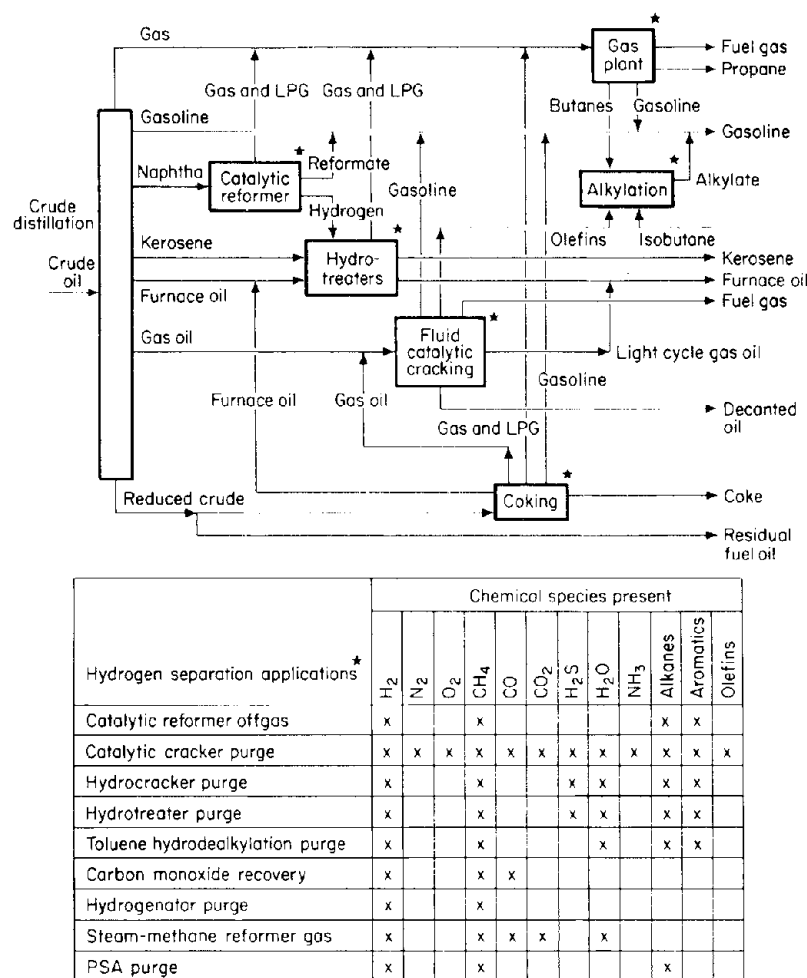


FIGURE 13 Hydrogen separation applications in the refinery. [From S. Leeper *et al.* (1984). Report No. EGG-2282, EG&G Idaho, Inc. (Report to U.S. Department of Energy), and D. L. MacLean *et al.* (1983). *Hydrocarbon Processing* 62, 47–51.]

other hydrogen-rich streams in the chemical process industry.

2. Air Separation

The products of air separation are oxygen and nitrogen at various purities. Oxygen-enriched air containing 30–40% O₂ can be used to increase the efficiency of combustion and other oxidation processes. Biochemical processes and organic chemical oxidations also benefit from the use of oxygen-enriched air to increase reaction rates and yields; an advantage of using membrane-processed air is that airborne impurities are thoroughly removed, thus reducing contamination. Nitrogen at 90–99% purity provides an inert atmosphere useful for various purposes: blanketing fuel storage tanks and pipelines to minimize fire hazards; reducing oxidation during annealing, sintering, and other

metal working operations; and retarding spoilage of foods during transport and storage.

Until the commercialization of membrane-based air separation systems in the mid-1980s, oxygen and nitrogen have traditionally been supplied in bulk by cryogenic systems via fractional distillation of liquified air, and by PSA. In many cases those conventional technologies remain competitive, especially for large-scale installations where their cost per unit capacity is favorable. By comparison, membrane systems require lower capital investment and operate at high efficiency over a wider range of reduced capacities. There are no expenses associated with storage and transportation of liquified gases. Mechanical problems or performance degradation due to inadequate feed air pretreatment are less likely to develop than in PSA. Recognizing the advantages of membrane systems, major gas producers have collaborated with suppliers of

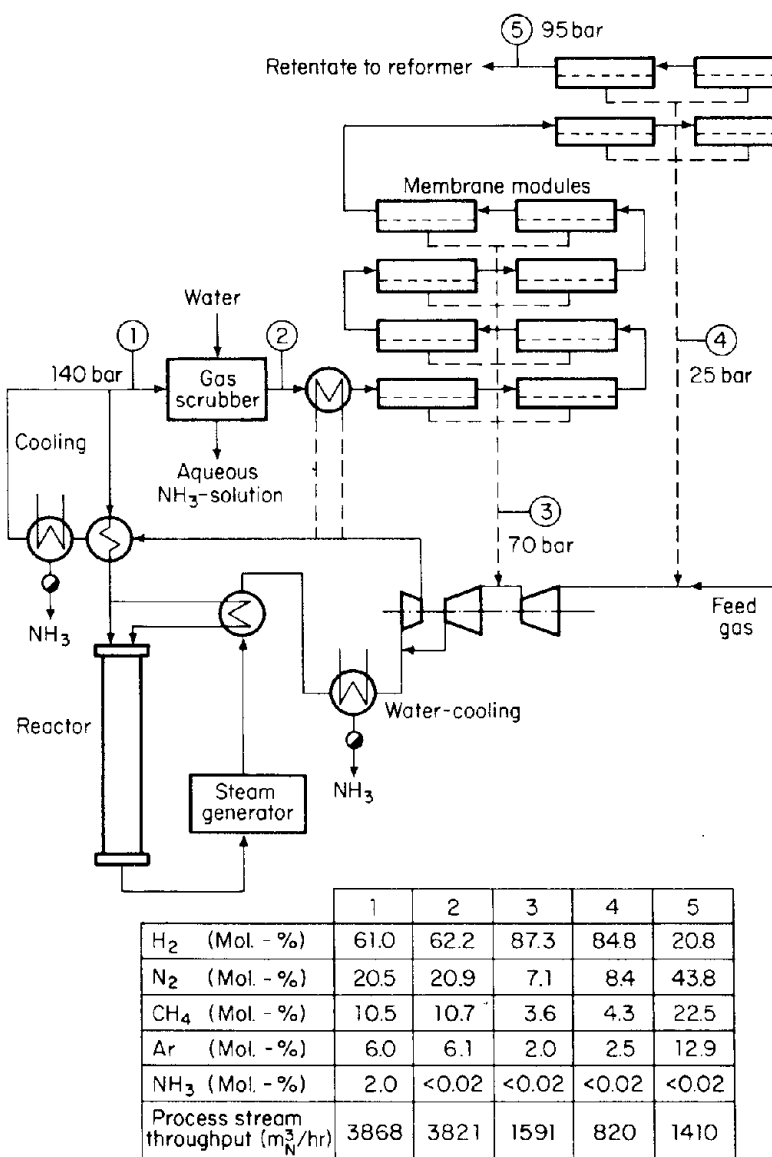


FIGURE 14 Hydrogen recovery scheme in ammonia synthesis. [From R. Rautenbach and R. Albrecht (1985). *Chem.-Ing.-Tech.* **57**, 119–130.]

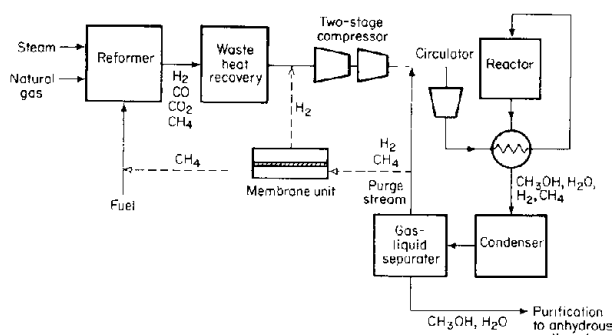


FIGURE 15 Hydrogen recovery scheme in methanol synthesis.

membranes to offer membrane systems for smaller users, and have successfully engineered hybrid cryogenic and PSA systems, which include membranes as an integral component. Figure 16 illustrates the complementary roles of different nitrogen- and oxygen-producing technologies for different purity and capacity requirements.

A commercial nitrogen enrichment system is illustrated in Fig. 17. Hollow-fiber membrane modules are connected to a compressed air feed at 70–150 psi. The feed in usually to the bore side of the hollow fibers. Oxygen (and water vapor that may be present) permeate out of the fiber into the shell and exit at low pressure. Dry, nitrogen-enriched air

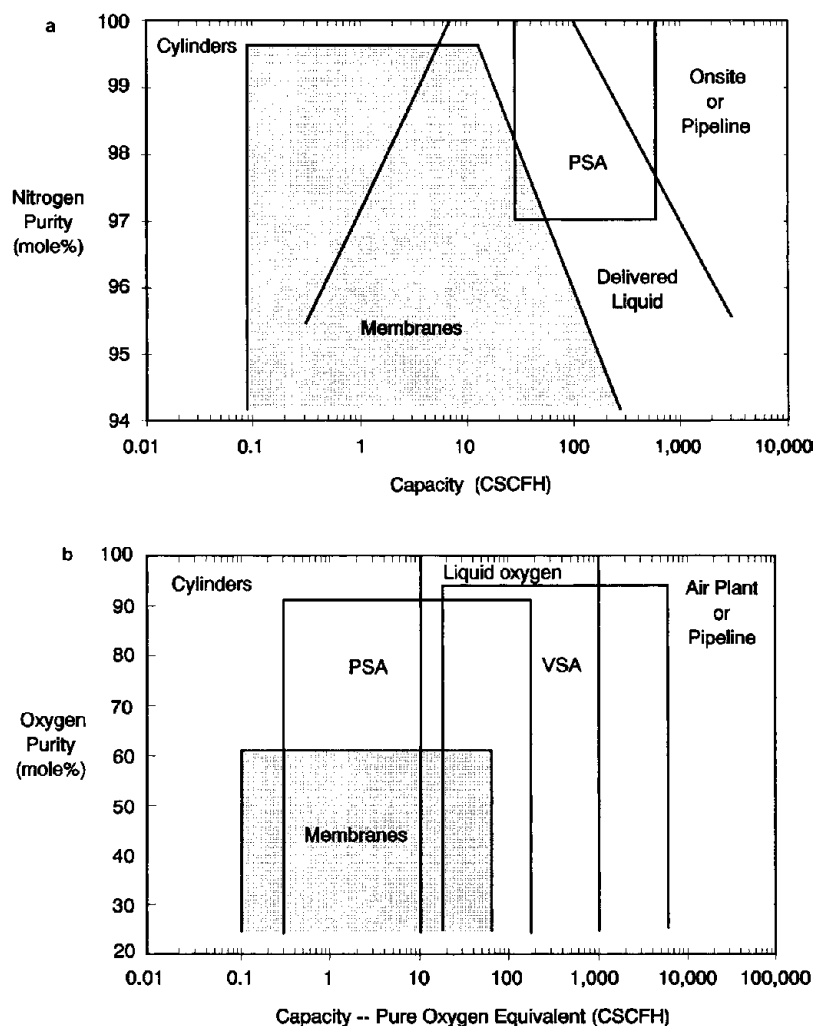


FIGURE 16 “Technology maps” showing feasibility of various air separation technologies for (a) nitrogen generation and (b) oxygen enrichment, using membranes compared to competitive technologies.



FIGURE 17 Compact nitrogen enrichment membrane system. (Courtesy Air Liquide.)

left inside the fibers is collected at a pressure slightly below that of the feed. Increasing the residence time of the feed air through the hollow fiber bundle reduces product output, but also results in more thorough removal of oxygen. By accepting a relatively low yield (<40% recovery), up to 99.5%-pure nitrogen can be obtained. The membranes available today are between three- to seven fold more permeable to oxygen than to nitrogen. With the most selective of these membranes, both oxygen-enriched air and high-purity nitrogen can be produced from the same system. However, the purity of oxygen is limited to 45–50% with present membranes because of the significant quantity of nitrogen entering the permeate stream.

Considerable effort is being devoted to developing new polymeric membrane materials. A special type of oxygen-enrichment membrane has also been explored, which consists of a solvent immobilized within a microporous solid support (Fig. 7D). Dissolved in the liquid is a carrier

compound that preferentially and reversibly complexes with oxygen. By exposing this liquid membrane (*q.v.*) to air on one side and a vacuum on the other side, the carriers capture oxygen from the air and release it on the evacuated side. The result is substantial enrichment of oxygen. However, many technical problems remain regarding the useful life of the carrier and of the membrane structure itself; commercial realization of this approach is unlikely. Also, as noted in the discussion of Fig. 7, membranes made from inorganic materials, e.g., metals or ceramics, sometimes formed as composites are under active study. Although not commercially significant, these membranes offer the potential of much improved chemical and thermal resistance, and might foster new industrial application possibilities.

3. Acid Gas Removal

Carbon dioxide, hydrogen sulfide, hydrogen chloride, sulfur dioxide, and some oxides of nitrogen are collectively referred to as acid gases. They are responsible for the “sour” nature of fuel gases from various sources. Carbon dioxide separation was the first large-scale applications developed; it remains one of the most important because of the abundance of carbon dioxide in gas mixtures. Another factor is that membranes exhibiting high permeability toward polar gases were relatively easy to develop based on desalination membrane technology already available.

Large quantities of carbon dioxide have been used in times of high energy costs in conjunction with enhanced oil recovery (EOR). This practice of injecting the gas at high pressure into geological formations to increase oil and natural gas yield, although highly effective, also in-

creases the concentration of carbon dioxide in the natural gas produced and thus reduces its energy value. Moreover, the cost of EOR would be prohibitively high unless the carbon dioxide is reused. Membranes have been used successfully at production wells to separate the carbon dioxide for reinjection while delivering a purified methane stream. Figure 18 shows a multistage process for reducing the carbon dioxide level of natural gas from 7 to 2% while achieving almost 95% methane recovery. Using carbon dioxide, permselective membranes such as cellulose acetate or various polyimides, the product natural gas stream remains at high pressure and requires little recompression for further processing. Such systems have proved to be much more economical to install and operate than diethylamine absorption, the prevailing method of gas treating.

Depending on source, geographic location, and the extent of extraction, the acid gas content of fuel gases often exceeds pipeline specifications. Certain natural gases and landfill gases can contain up to 50% carbon dioxide. Bulk removal of both carbon dioxide and hydrogen sulfide from such sources, i.e., the process of “sweetening,” not only improves the fuel value of the gas, but also helps reduce corrosion of pipelines and transmission equipment. Membranes are suitable for this application especially where the scale is relatively small and the economics favor scalable membrane systems.

Sulfur dioxide is a common pollutant found in coal-fired facilities. Various membrane permeation schemes have been proposed but few are competitive with wet scrubbing. More recently, however, bipolar membrane technology (*q.v.*) has been successfully used to recycle the scrubbing effluent and convert the sulfur into sulfuric acid.

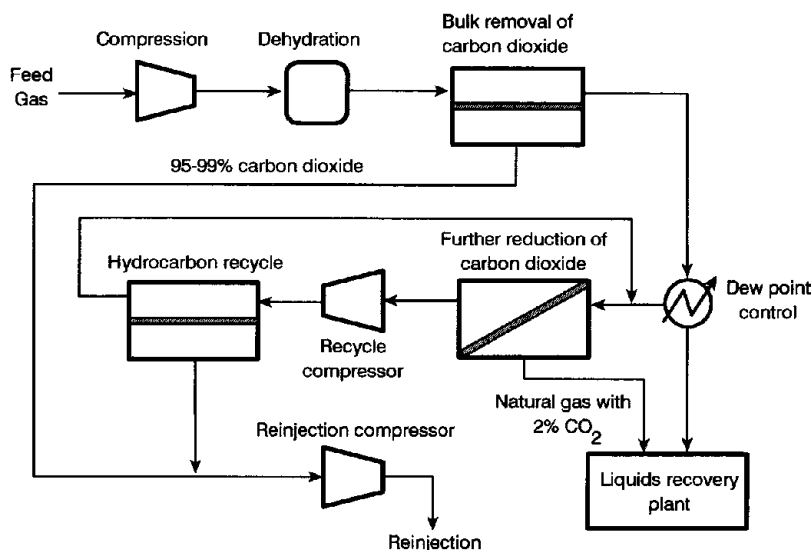


FIGURE 18 Multistage carbon dioxide recovery from natural gas in enhanced oil recovery operation.

4. Water Vapor Separation

Increasing environmental awareness and rising costs of energy and chemical supplies have helped spur interest in membrane processes as a means of recovering those resources. Although vapor separation is closely related to gas permeation in mechanism, the presence of condensable components permits unique process designs and opportunities for energy recovery.

Water vapor separation from air is an example where significant energy recovery is possible. Many industries, including pulp-and-paper, textile, and food processing, use large amounts of energy to dry their products. The water vapor in the exhaust air from the dryer carries with it the latent heat of vaporization. Reclaiming this latent heat could substantially reduce the net thermal energy requirement of the drying process. Heat exchange between incoming and exhaust streams is inefficient because only sensible heat is recovered, and because the temperature differential is usually small. Condensing the water vapor by recompressing the moist exhaust air would release the latent heat, but this approach is also inefficient because most of the energy is expended in compressing the major component—air. A membrane-aided vapor-compression process shown in Fig. 19 could be a more efficient alternative. Moisture from the dryer exhaust would pass through a hydrophilic membrane and be compressed. The latent heat liberated can be used to preheat the feed air for the dryer. It has been estimated that more than half of the energy used in a paper drying machine can be reclaimed in this way. Economic feasibility of this scheme hinges on the relative costs of energy and that of the membrane system.

Already commercialized is an innovative membrane module capable of in-line drying of air. Moist air is fed to one side of an inherently water-permeable membrane, which has a low density of surface pores. While water vapor diffuses through the membrane preferentially, a small

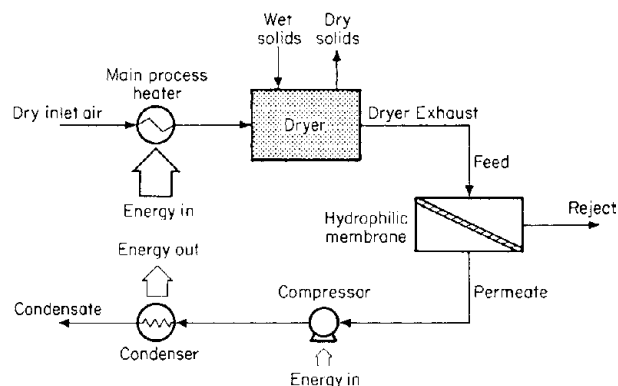


FIGURE 19 Concept of membrane-assisted recompression for energy recovery from drying operations. (Bend Research, Inc.)

amount of air also leaks through the pores and carries the water vapor away from the membrane. In this way, the accumulated water vapor in the membrane sublayer is continually swept away, thereby preventing condensation and loss of water removal capability. Micropores normally considered defects in the membrane are in fact a necessary feature in this membrane design. Very low dew points can be reached with this membrane unit with no additional power input and no moving parts. The inherent reliability and simplicity of this product makes it attractive for instrumentation applications and at locations that are difficult to access.

5. Organic Vapor Separations

Organic vapor separation from air is a means of controlling pollution and recovering valuable resources. An indication of the environmental problem is the more than 30 million tons per year of volatile organics emitted in 1975 from all stationary sources in the United States. These included petroleum refineries, chemical plants, and defective solvent storage and conveyance facilities. Much of the vapors emitted—hydrocarbons, chlorinated solvents, alcohols, and ketones—are potentially recoverable with membranes highly permeable to organics and relatively impermeable to air. A scheme for treating a solvent-laden air stream from a drying oven is shown in Fig. 20, where liquid solvent is recovered by compressing the permeate stream, and hot air is recycled to the drying oven.

In gas- or vapor-phase chemical reactors, the product stream typically contains some residual reactants and one or more inert gases. At the end of the reaction cycle the product is recovered—for example, by condensation—and the remaining gases are vented to the atmosphere or flared. Significant quantities of reactants can be lost in this way, some adding to the environmental burden of the process. Membrane systems have been designed to address both issues. By selecting a membrane more permeable to the reactant than to the inert gases, a highly concentrated reactant stream can be collected as permeate and recycled

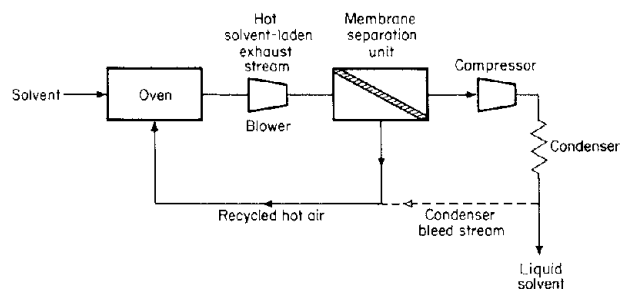
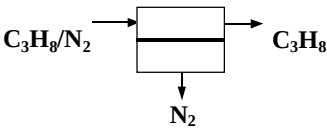
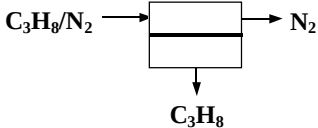


FIGURE 20 A membrane system for solvent recovery and waste heat capture. (Membrane Technology and Research, Inc.)

TABLE VI Guidelines for Membrane Selection in Gas-Vapor Separations (Baker *et al.*, 1998)

	Gas-selective membrane	Vapor-selective membrane
Configuration		
Membrane	Rigid, glassy, amorphous (e.g., polysulfone)	Soft, rubbery (e.g., silicone rubber)
Typical nitrogen permeation flux (10^{-6} cm ³ (STP)/cm ² sec cm Hg)	1–10	100–1000
Typical separations	H ₂ /N ₂ , O ₂ /N ₂ , CO ₂ /CH ₄	Volatile organic compounds/air, C ₃ H ₈ /N ₂

to the reactor. Conversely, using a membrane more permeable to the inert gas can yield a reactant stream at pressure, reducing or eliminating the need for recompression during recycle. Some general guidelines required for practical application of these two strategies are given in Table VI (Baker *et al.*, 1998). Attractive applications have been implemented for recovering unreacted monomers from polymerization plants—for example, vinyl chloride monomer from a polyvinyl chloride plant, as shown in Fig. 21, and olefin mixtures from polyethylene and polypropylene plants Fig. 22. (Baker *et al.*, 2000). Both of these practical cases rely upon existing materials that preferentially pass the organic vapor, and no real examples of commercially viable gas- vs vapor-selective membranes have been reported yet.

IV. VAPOR–LIQUID SEPARATIONS

A. Pervaporation

When a liquid mixture evaporates freely, the vapor-phase composition is governed solely by thermodynamic equilibrium.

However, if the liquid is bound by a nonporous membrane, then the rate of vaporization of each component is limited by the permeability of the membrane. This membrane-mediated evaporation process is referred to as pervaporation. By providing a partial vacuum or by circulating a noncondensing sweep gas on the downstream side of the membrane, as shown in Fig. 23, permeate vapor is continuously withdrawn and optionally condensed as liquid. Here, the composition of the permeate is governed by both the feed composition and the permselectivity of the membrane. Pervaporation may therefore be thought of as membrane-mediated evaporation.

Pervaporation is superior to reverse osmosis for separating organic–organic mixtures because of the very high osmotic pressures associated with such systems, for which the net driving force [$\Delta p - \Delta \pi$] is unacceptably low for any realistic pumping pressure. In contrast, pervaporation may be operated using relatively modest pressures for creating a vacuum on the downstream side of the membrane, relying on the large drop in activity between the liquid and vapor phases of the permeant as the primary driving force.

The pervaporation separation factor, β_{pervap} , simply equals the product of the evaporation (distillation)

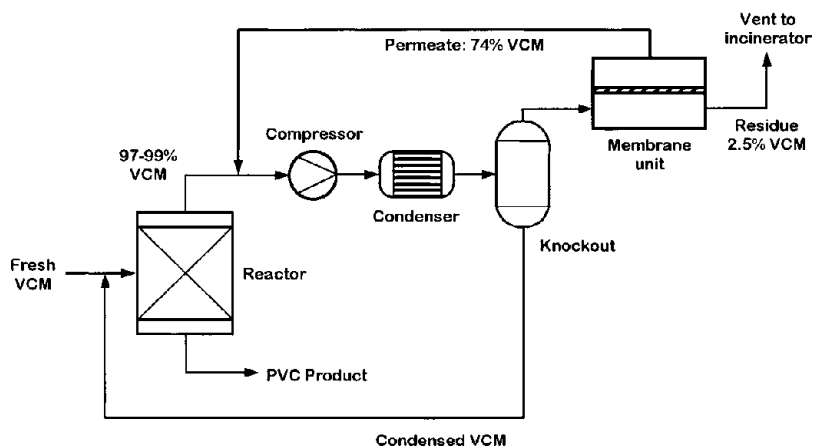


FIGURE 21 Membrane-based vinyl chloride monomer (VCM) recovery system in a PVC polymerization plant. (Membrane Technology and Research, Inc.) [From Baker, R. W., *et al.* (2000, December). *Chem. Eng. Prog.* pp. 51–57.]

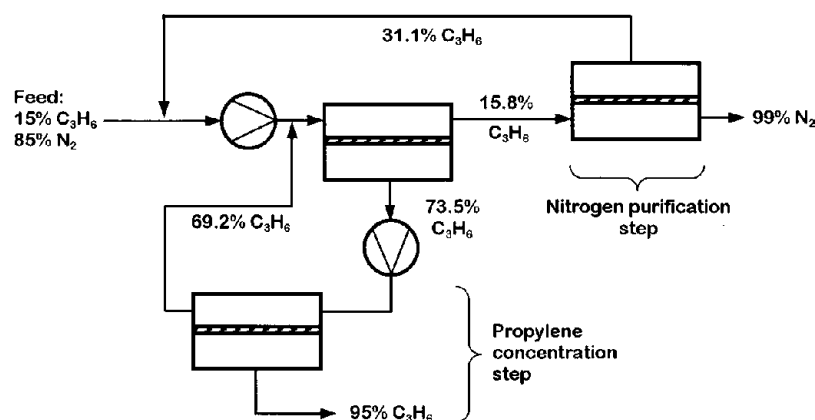


FIGURE 22 Two-step, two-stage membrane system to produce pure nitrogen and a concentrated propylene stream from a dilute feed gas. (Membrane Technology and Research, Inc.) [Baker, R. W., *et al.* (1998). *J. Membrane Sci.* 159, 55–62.]

separation factor, β_{evap} , and the membrane-moderated factor, β_{mem} , viz.,

$$\beta_{\text{pervap}} = \beta_{\text{evap}} \cdot \beta_{\text{mem}} \quad (15)$$

This form stresses that part of the separation in the pervaporation process occurs *independent of the presence of the membrane*, β_{evap} . Equation (15) also stresses that part of the separation relies *strictly* on the identity of the membrane material being used, β_{mem} . In this context, the membrane is seen as separating a hypothetical vapor feed (in equilibrium with the actual liquid feed) and the downstream vapor permeate product.

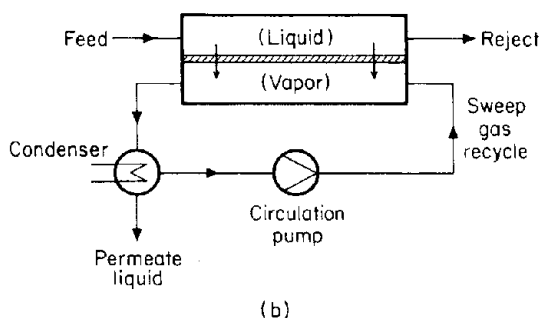
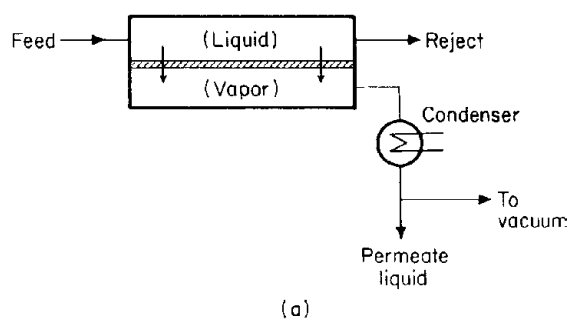


FIGURE 23 Pervaporation process concepts: (a) partial vacuum operation and (b) permeate-side sweep stream operation.

At low permeate pressures typical of ideal pervaporation, the same form for membrane selectivity applies as for gas or vapor separation of components A and B, which will be discussed later. This simple selectivity equals the inverse ratio of the resistances of the membrane to permeation of components A and B [Ω_B / Ω_A from Eq. (1)]. This ratio can be shown to simply equal to the ratio of permeabilities of the two components in the material comprising the selective layer of the membrane.

$$\beta_{\text{mem}} = \frac{P_A}{P_B} = \alpha_{\text{mem}} \quad (16)$$

Factors governing P_A and P_B are understood, so the key scientific issues in pervaporation materials selection are similar to those in gas separation.

Significant opportunities exist for pervaporation in niches where distillation has a weakness, such as with azeotropes and close-boiling point organic–organic mixtures, whose composition cannot be changed by conventional distillation unless the thermodynamic equilibrium is shifted by additional components (as in extractive distillation). When size differences between liquid components are about an Ångström or less, fine size discrimination again requires careful consideration, as in the case for gases discussed above. Especially when solubility selectivity offers little advantages, such as with isomers like *o*- and *p*-xylene, or *n*- and *i*-butane, control of mobility selectivity offers special opportunities. Efficient module designs are critical to ensure adequate heat transfer and solvent resistance, and these aspects still require significant development.

1. Alcohol Dehydration

Ethanol–water separation was the first industrial-scale application of pervaporation. It remains the dominant use of

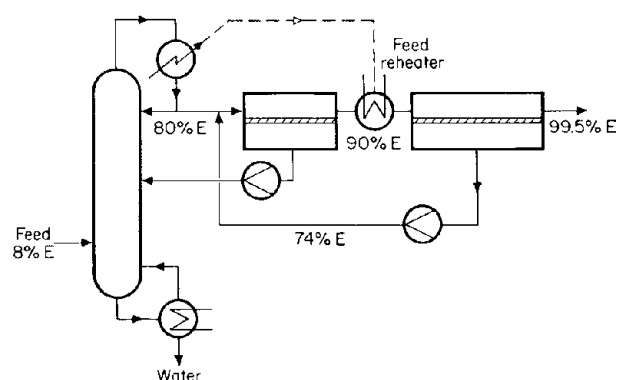


FIGURE 24 Hybrid distillation–pervaporation process for fermentation ethanol production (E, ethanol).

the technology since the beginning of the 1990s. Conventionally, anhydrous ethanol is produced by distilling dilute alcohol from about 10% (e.g., that from biomass fermentation) to about 90%, then further dehydrating to 99+% ethanol by means of azeotropic distillation or extractive distillation. Dehydration requires additional columns, besides the main rectification tower and the use of entrainers to break the azeotrope. Although distillation enriches ethanol efficiently from low to moderately high concentrations (ca. 80%) in a small number of equilibrium stages, it becomes increasingly energy-intensive as the azeotropic point is approached. The hybrid system shown conceptually in Fig. 24 replaces the energy- and capital-intensive

portion of the conventional process with a pervaporation unit featuring water-selective membranes. Overhead vapor from the distillation section is condensed and fed to the membrane unit at a temperature slightly below the boiling point. Water pervaporates preferentially, leaving purified ethanol on the feed side. To supply the latent heat of vaporization, a pervaporation system is designed with reheaters between succeeding membrane stages. Industrially, pervaporation technology for ethanol/water separation has matured rapidly. Modern pervaporation plants have capacities reaching several thousand tons per year.

Separation of isopropanol (IPA) and water by pervaporation has also reached production scale. Much of the current capacity is devoted to azeotrope breaking and dehydration during IPA synthesis. Recently, anhydrous isopropanol has become a preferred drying solvent in the semiconductor industry, where chip wafers are first washed with ultrapure water, then rinsed with the alcohol to promote uniform drying. The water-laden isopropanol generated can be conveniently reused after dehydration by pervaporation. Unlike with pressure-driven membrane processes such as RO or UF, particulates and nonvolatile substances such as salts are not carried over during pervaporation. This helps maintain the effectiveness of contamination control.

Pervaporation technology has matured considerably over the past two decades. Increasing numbers of applications have been identified, such as those listed in Table VII.

TABLE VII Products Separated or Purified by Pervaporation (Source: Sulzer Chemtech Ltd., Winterthur, Switzerland)

Alcohols		Esters	
Methanol	CH ₄ O	Methyl acetate (MeAc)	C ₃ H ₆ O ₂
Ethanol	C ₂ H ₆ O	Ethyl acetate (EtAc)	C ₄ H ₈ O ₂
Propanol (both isomers)	C ₃ H ₈ O	Butyl acetate (BuAc)	C ₆ H ₁₂ O ₂
Butanol (all isomers)	C ₄ H ₁₀ O	Ethers	
Pentanol (all isomers)	C ₅ H ₁₂ O	Methyl <i>tert</i> -butyl ether (MTBE)	C ₅ H ₁₂ O
Cyclohexanol	C ₆ H ₁₂ O	Ethyl <i>tert</i> -butyl ether (ETBE)	C ₆ H ₁₄ O
Benzyl alcohol	C ₇ H ₈ O	Di-isopropyl ether (DIPE)	C ₆ H ₁₄ O
Ketones		Tetrahydro furan (THF)	C ₄ H ₈ O
Acetone	C ₃ H ₆ O	Dioxane	C ₄ H ₈ O ₂
Butanone (MEK)	C ₄ H ₈ O	Organic acids	
Methyl isobutyl ketone (MIBK)	C ₆ H ₁₂ O	Acetic acid	C ₂ H ₄ O ₂
Aromatics		Nitriles	
Benzene	C ₆ H ₆	Acetonitrile	C ₂ H ₃ N
Toluene	C ₇ H ₈	Aliphatics	From C ₃ to C ₈
Phenol	C ₅ H ₆ O	Chlorinated hydrocarbons	
Amines		Dichloro methane	CH ₂ Cl ₂
Triethylamine	C ₆ H ₁₅ N	Perchloroethylene	C ₂ Cl ₄
Pyridine	C ₆ H ₅ N		
Aniline	C ₆ H ₇ N		

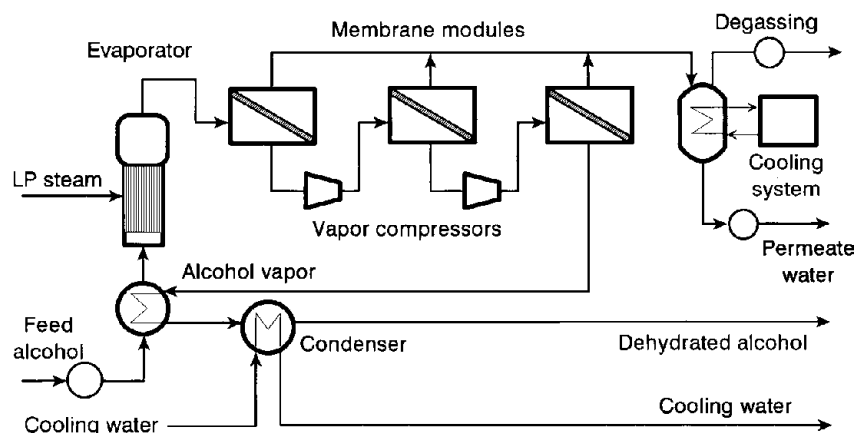


FIGURE 25 Medium-scale (30,000 liters/day) ethanol dehydration system by vapor permeation (Lurgi GmbH).

2. Vapor Permeation

A variant of the pervaporation process was commercialized very recently for ethanol/water separation. The ethanol feed solution is vaporized before delivery to the membrane unit; a vacuum is provided to remove the permeate. This process thus has as much in common with gas separation as it does pervaporation. In contrast with pervaporation, however, uniformly high temperatures can be maintained throughout the membrane unit without interstage reheating because the enthalpy of vaporization need not be supplied adiabatically from the feed stream. Since permeation flux varies directly with temperature, operation under essentially isothermal conditions results in a higher overall water removal rate. Consequently, fewer stages are needed to reach a given ethanol purity. Figure 25 shows the process flow diagram of a large vapor permeation plant in operation, with a capacity of dehydrating 30,000 L/day of 94% ethanol to 99.9%.

3. Trace Organic Compound Removal from Water

In the United States the emission of volatile organic compounds (VOCs) has been estimated to be as high as 5 million metric tons annually. These pollutants are found mostly in wastewater or contaminated groundwater. Pervaporation has been shown in various pilot-scale studies to be well suited for purifying such wastewaters. A generic design of such a system is shown in Fig. 26. Particularly attractive are cases in which the value of the recovered solvent is high (e.g., halogenated hydrocarbons or alkyl esters) such that the payback period is relatively short, or the hazard present by the compounds warrants even costly remedial and restorative measures. Often, the most cost-effective way of deploying pervaporation systems is to combine it with a secondary method of clean up, such as activated carbon absorption. The pervaporation system

would reduce initial high concentrations of VOCs, while carbon absorption further reduces those contaminants to trace levels.

Besides wastewater treatment, pervaporation systems have also been tested on a development scale for continually removing volatile organic products (e.g., ethanol, volatile acids) from fermentation broths.

4. Production of Organic Acid Esters

Fermentation-derived organic acids and their esters are potentially important chemical feedstocks for polymers and specialty polymers, but most significantly as alternative solvents for industrial and consumer applications. For example, lactate esters are derived from renewable carbohydrate raw materials such as cornstarch. They exhibit much lower toxicity compared with halogenated hydrocarbons and ethylene glycol ethers and are environmentally benign. Some studies suggested that lactate ester solvents have the potential of replacing petroleum-based solvents

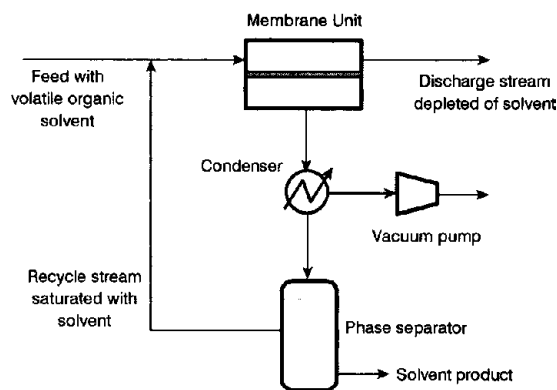


FIGURE 26 Removal of trace volatile organic compounds from wastewater streams by pervaporation. (Membrane Technology and Research, Inc.)

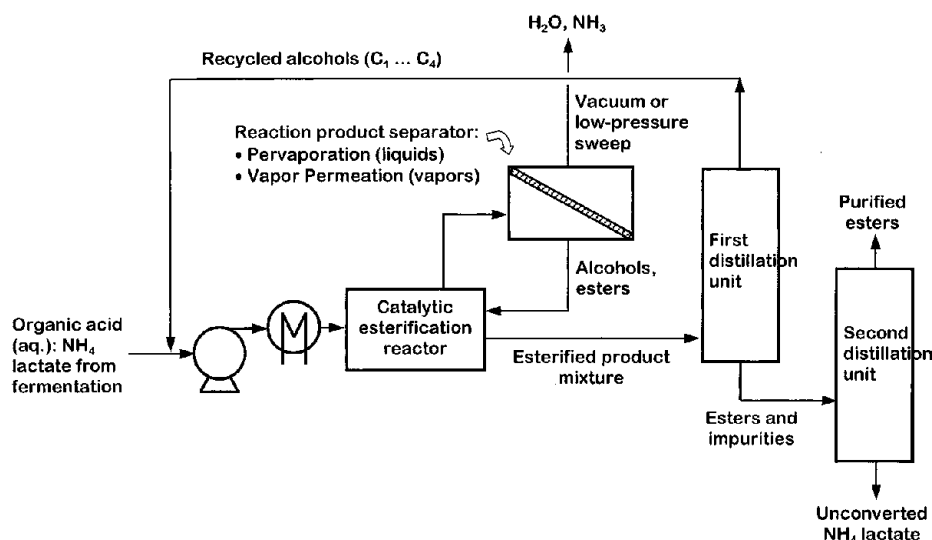
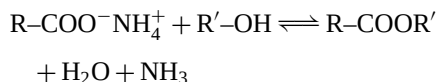


FIGURE 27 Pervaporation-assisted organic acid ester production (Argonne National Laboratory). [U.S. Patent 5,723,639.]

if the production cost can be reduced significantly below \$1/lb, and if adequate supply and distribution systems are established (Datta and Tsai, 1998).

In a conventional process, ammonium salts of organic acids are derived first by fermentation and then catalytically cracked to yield the acid. The acid then reacts with C_1 – C_4 alcohols to form the ester. The reaction products include the acid ester, unreacted alcohol, water, and ammonia:



The water and ammonia formed inhibit the forward reaction; their buildup in the reactor limits the yield of the ester product.

To improve process economics, an integrated process shown conceptually in Fig. 27 has been proposed. A pervaporation subsystem is equipped with a membrane selectively permeable to water and ammonia, but rejects ethanol and ethyl lactate. The retentate stream carrying these reactants may be returned to the reactor to help drive the reaction toward completion.

B. Membrane Distillation

Membrane distillation involves partially evaporating a solution through a microporous membrane that is vapor-permeable but liquid-repellent. The membrane has no permselectivity, but provides a stable liquid–gas–liquid interface for vapor transfer. As shown in Fig. 28, the membrane separates a heated feed solution from a cooler product solution. Since the vapor pressure of solvent in the feed solution is higher than that in the product solution, solvent

vapor flows through the air-filled pores of the membrane and condenses on the product side as “distillate.” Relatively pure solvent can be recovered, provided that the solutes have low volatility. In all cases, vapor–liquid equilibrium determines the degree of separation achievable.

Membrane distillation was first applied to seawater desalination in the early 1980s, using microporous polytetrafluoroethylene membranes as the hydrophobic barrier. High fluxes and good product water quality could be obtained over a wide feed concentration range. Fouling and scaling problems were less acute compared with that in reverse osmosis. Development and process improvements continued into the 1990s (Balaban, 1991). Nevertheless, the process was not economically competitive with reverse osmosis except in special situations because the membrane cost and the costs associated with capturing and transferring “waste heat” supplies are often too high relative to the value of the desalted water. More recently, membrane distillation has been targeted for higher-valued applications such as ultrapure water production and food concentration.

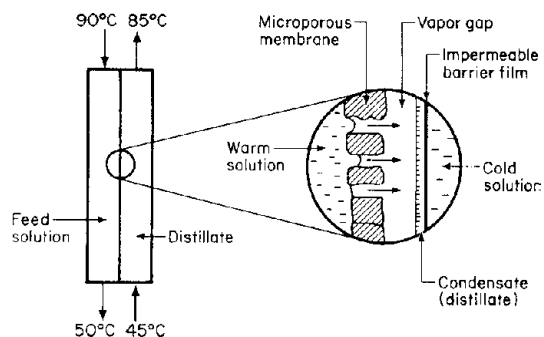


FIGURE 28 Mechanism of membrane distillation.

C. Osmotic Distillation

Osmotic distillation also removes the solvent from a solution through a microporous membrane that is not wetted by the liquid phase. Unlike membrane distillation, which uses a thermal gradient to manipulate the activity of the solvent on the two sides of the membrane, an activity gradient in osmotic distillation is created by using a brine or other concentrated solution in which the activity of the solvent is depressed. Solvent transport occurs at a rate proportional to the local activity gradient. Since the process operates essentially isothermally, heat-sensitive solutions may be concentrated quickly without an adverse effect. Commercially, osmotic distillation has been used to de-water fruit juices and liquid foods. In principle, pharmaceuticals and other delicate solutes may also be processed in this way.

D. Vapor-Arbitrated Pervaporation

In certain cases it is desirable to selectively remove a volatile solute from a solution that contains other, less volatile, solutes as well as the solvent. Some examples are the reduction of ethanol content from alcoholic beverages or from dilute alcoholic extracts of aromatic flavors and fragrances from plant sources such as fruits or flowers. Conventional pervaporation would facilitate removal of water from such mixtures while retaining ethanol and the higher molecular weight organics that comprise the characteristic aroma and flavor profile of the products of interest. On the other hand, membrane distillation or osmotic distillation cannot retain the volatile components at all.

A process referred to as vapor-arbitrated pervaporation addresses these issues by manipulating the transmembrane activity gradients of water and ethanol in a pervaporation system. Using a permeate side sweep stream that contains water vapor at a partial pressure corresponding to the activity of water on the feed side, permeation of water is halted while ethanol continues to diffuse through the membrane into the sweep stream and is removed. In this way, the native permselectivity of the membrane system can be altered in a controlled fashion to extract one or more volatile components from a solution.

The concept of this process has been demonstrated for lowering or increasing the alcohol content of distilled spirits by using water vapor or ethanol vapor in the sweep stream, respectively. In either case, this vapor arbitration action combined with the inherent selectivity of the membrane resulted in virtually complete preservation of the subtle character of the beverage, but in a more concentrated form due to the lower net volume of the retentate product (Lee, 1993). Similar results may be anticipated in other volume reduction applications involving high-value volatile feedstocks.

One approach to delivering increased performance in a membrane process is to complement one separation mechanism with another. Vapor-arbitrated pervaporation is an example of this strategy. In bioseparations, as will be covered in a later section, a similar integration of several process enhancements in High-Performance Tangential Flow Filtration is responsible for dramatic improvement in separation efficiency of protein mixtures once considered unachievable by means of conventional ultrafiltration.

V. LIQUID SEPARATIONS

Membrane processes have been applied successfully to a wide variety of liquid separations. Table VIII lists a number of typical applications by industry and by technology. In the following sections, the function and applications of each process are illustrated by commercialized examples. The outlook of each technology segment is also discussed.

A. Reverse Osmosis

RO occurs when a solution is pressurized against a solvent-selective membrane, and the applied pressure exceeds the osmotic pressure difference across the membrane. Water is the solvent in most existing reverse osmosis applications; the solutes may be salts or organic compounds.

Reverse osmosis for desalting seawater and brackish water was the first industrial-scale application of modern membrane technology. The principles and practice of RO technology are well established, with a worldwide desalination capacity reaching 6.8 billion gallons of water per day at the end of 1999. Several factors contributed to the success of reverse osmosis for desalination: the process is more energy-efficient than distillation, high-flux membranes with good salt rejection have become more durable, lower cost commodity products. Modern RO plants are capable of producing potable water at less than \$1/m³ including all capital and operating costs.

Seawater contains about 3.5 wt % of total dissolved solids (TDS) in most locations of the world. Typical RO systems operate between 50 and 70 bars, and require less than 10 kWh in energy to produce one cubic meter of potable water with less than 0.05 wt % TDS. This is substantially lower than the 15–16 kWh/m³ required for multistage flash distillation technology. Although thermal desalination is well established and reliable, the energy advantage of reverse osmosis favors the overall economics of membrane systems as large as 75,000 m³/day in capacity. Figure 29 shows a seawater reverse osmosis desalination facility. Over the last two decades, reverse osmosis has captured an increasing share of the desalination market previously dominated by distillation—even

TABLE VIII Liquid Separations^a

Industry and applications		Liquid-to-liquid								Liquid-to-vapor	
		Pressure driven			Concentration driven				Electrically driven	Pressure and concentration driven, PV	Temperature driven, MD
		RO	UF	MF	D	DD	MSX	CT	ED	EDR	
Desalination	Potable water production										
	Seawater	*									
	Brackish water	*							*	*	*
	Municipal wastewater reclamation	*	*								
Utilities and power generation	Ultrapure water production	*	*	*					*	*	*
	Boiler feedwater production	*							*	*	
	Cooling tower blowdown recycle	*									
Metals and metal finishing	Hydrometallurgy							*	*		
	Mining effluent treatment	*					*	*	*		
	Plating rinse water reuse and recovery of metals	*	*			*		*	*		
	Oil–water separation		*								
	Electrocoat paint recovery		*								
Food processing	Dairy processing	*	*	*					*		
	Sweeteners concentration	*									
	Juice and beverage processing	*	*	*	*				*		*
	Protein recovery, concentration, and purification		*	*		*	*		*		*
Textiles	Dyeing and finishing, chemical recovery, water reuse	*	*								
Pulp and paper	Black-liquor concentration		*								
	Effluent disposal and water reuse	*	*								
Chemical process industries	Process water production, reuse	*	*								
	Effluent disposal and water reuse	*	*						*	*	
	Chlor-alkali production								*		
	Electrochemical synthesis								*		
	Water–organic liquid separation	*									*
	Organic liquid mixture separation	*	*				*				*
Biotechnology/medicine	Fermentation products recovery and purification	*	*	*	*		*	*			*
	Cell harvesting, virus and antibody concentration		*	*							
	Protein desalting, concentration and fractionation		*						*		
	Blood processing, including artificial kidney		*	*	*						
Analytical	Isolation, concentration, and identification of solutes and particulates	*	*	*			*				

^a Key: RO, reverse osmosis; UF, ultrafiltration; MF, microfiltration; D, dialysis; DD, Donnan dialysis; MSX, membrane solvent extraction; CT, coupled transport; ED, electrodialysis; EDR, electrodialysis reversal; PV, pervaporation; MD, membrane distillation.

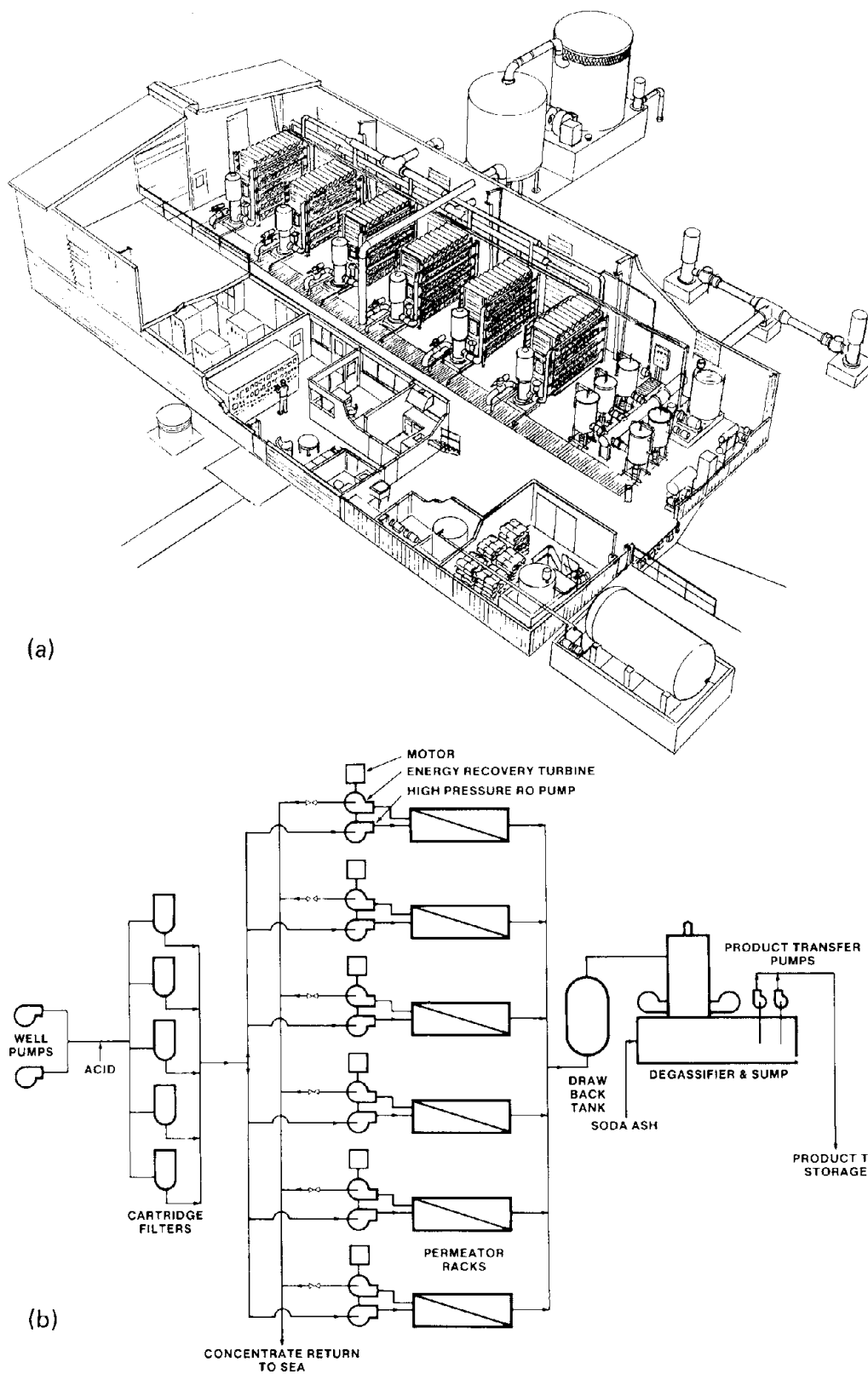


FIGURE 29 A large-scale reverse osmosis seawater desalination facility. (a) Plant layout schematic and (b) simplified process diagram. (Du Pont Company.)

for very large (up to 200,000 m³/day) desalination plants—a transition aided by sharply rising energy costs, and by high-performance, yet competitively priced, membrane systems. Energy-recovery turbines are used extensively in seawater RO systems to reclaim energy from the high-pressure brine stream. For the very high-salinity seawaters (>4.5% TDS) found in Middle East locations, desalination systems are designed to operate at about 80 bars.

Brackish waters contain between 0.05 and 1 wt % TDS. Their lower osmotic pressures allow reverse osmosis operation between 15 and 30 bar. Less expensive pressure equipment and energy consumption translate to more favorable water production economics than those for seawater desalination.

Reverse osmosis membranes can be divided into subclasses according to their solute/water selectivity and operating pressure regimes. Figure 30 shows a number of commercial membranes developed for seawater and brackish desalination, and for nanofiltration. These include cellulose ester and polyamide asymmetric membranes available since the 1960s, and high-performance composite membranes developed in the 1970s. Collectively, they make it possible to produce potable water from virtually all saline water sources.

A lingering limitation with the present generation of reverse osmosis membranes is their limited resistance to chemical attack. In particular, membranes derived from polyamides, polyureas, and other nitrogen-containing polymers are susceptible to oxidative degradation by chlorine—the most widely used disinfectant to pretreat feed waters. Dissolved oxygen can also damage reverse osmosis membranes when catalyzed by trace heavy metals. Successful development of oxidation-resistant membranes will help reduce the complexity and costs associated with the elaborate pretreatment now required.

Water supplied to industry has to meet stringent specifications. For example, process water for the chemical and biotechnology industries is routinely purified beyond potable water standards. Boiler feed water for steam generation must contain a minimum of silica. Reverse osmosis units designed specifically for these purposes are in widespread use today. For example, reverse osmosis/distillation hybrid systems have been designed to separate organic liquids. For semiconductor manufacture, reverse osmosis is combined with ultrafiltration, ion exchange, and activated carbon adsorption to produce the extremely clean water required.

Wastewater reclamation is a logical extension of desalination technology. Much of the membrane system design is common to both applications, and the membranes available for wastewater treatment are those originally developed for desalination. The first major project designed for

this purpose is Water Factory 21 located in Orange County on the California coast. In operation since 1976, the facility treats municipal wastewater by reverse osmosis and blends the product with water purified by carbon absorption and from deep wells. The combined stream, which meets drinking water standards, is reinjected into coastal aquifers to replenish local groundwater supplies and prevent seawater intrusion. At Yuma, Arizona, the world's second largest reverse osmosis plant, treats 275,000 m³/day of saline farmland drainage so that salinity requirements can be met for Colorado River water released to Mexico.

Liquefied and gasified coal have been considered as an alternative to petroleum for producing energy and as chemical feedstock. Both liquefaction and gasification generate large volumes of water from coal washing, slurry-ing, and the conversion process itself. These wastewaters are contaminated with salts, phenol, ammonia, hydrogen sulfide, and a complex mixture of other substances. Simultaneous removal of organics (up to 98%) and salts (between 80 and 95%) by reverse osmosis shows some promise.

Reverse osmosis also serves some of the waste management and resource recovery needs in the metals and metal finishing industry. Effluent streams from mining and plating operations containing heavy metals, acids, and other chemicals can be treated with reverse osmosis to recover both the metal as its salt, and purified water for reuse. For metal ion recovery from dilute solutions, however, reverse osmosis faces competition from conventional solvent extraction, membrane-based solvent extraction, and its variant, coupled transport (see Section V.F.3).

An estimated 10¹⁵ KJ are consumed annually in the United States for food processing, primarily in concentration and purification operations. Concentration by reverse osmosis is attractive because of its ability to remove water without adding heat, and is already used for concentrating sugar solutions, fruit and vegetable juices, and beverages while retaining salts and low-molecular-weight flavor components. Ambient temperature processing also helps preserve product quality. High concentrations are reached by using membranes with high rejections and operating at very high pressures (100 bar or above) so as to overcome the osmotic pressures associated with increasing sugar contents. Sometimes membranes with lower rejection are used to recover residual solute in the permeate, or at the final stage of concentration where the osmotic pressure is at its maximum. In these applications, reverse osmosis and nanofiltration membranes are often deployed together to balance productivity, product specification, and cost.

The United States textile industry consumes over 4 billion m³ of water annually. Much of the process water is discharged together with dyes and auxiliary chemicals, plus a loss of energy in the hot effluents. Reverse osmosis

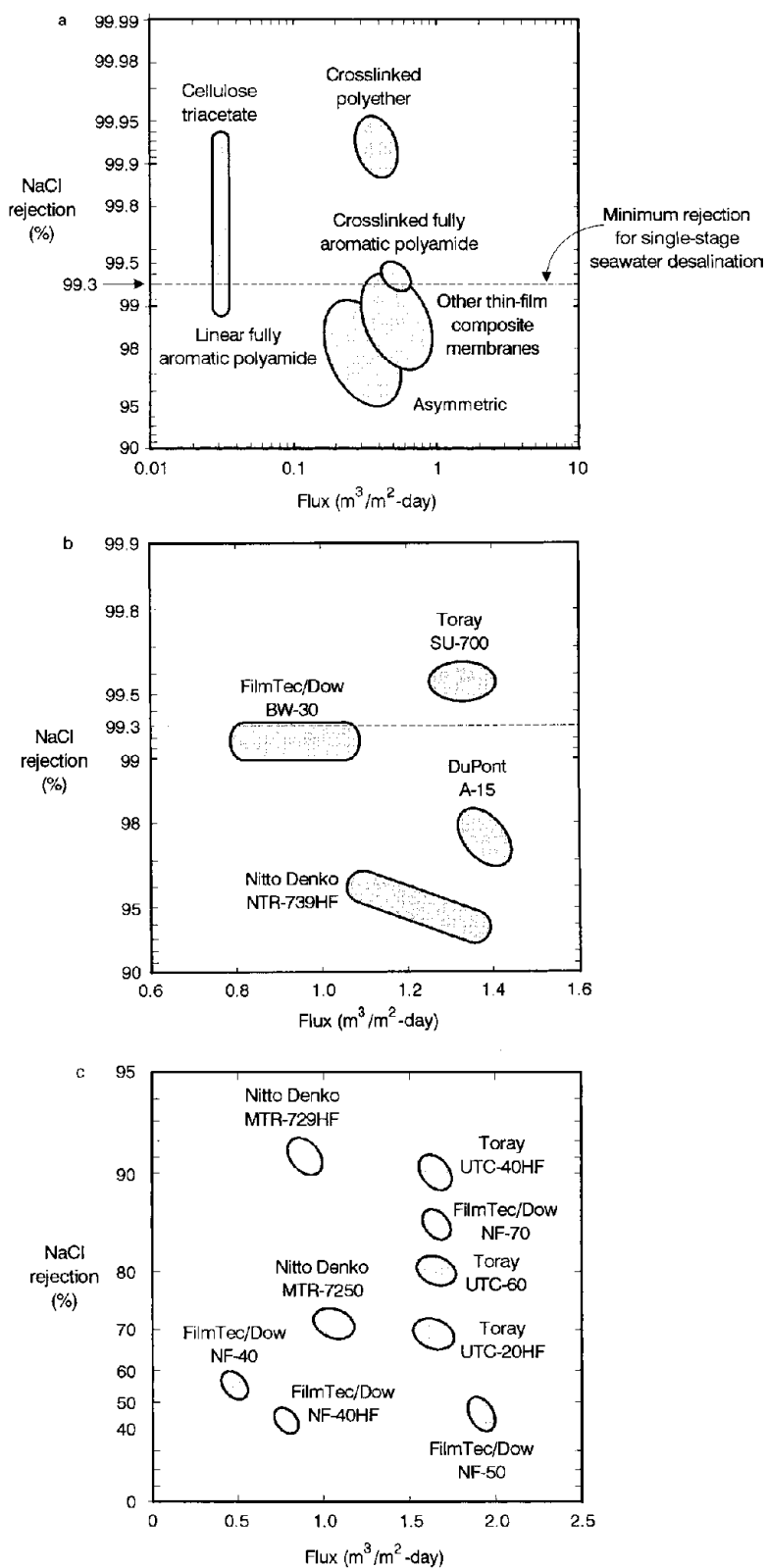


FIGURE 30 Performance of some commercial reverse osmosis membranes for (a) seawater desalination (test conditions: 56 bar; 25°C; 3.5% NaCl feed); (b) low-pressure desalination (15 bar; 25°C; 1500 mg/liter NaCl feed); and (c) ultralow-pressure nanofiltration applications (7.5 bar, 25°C; 500 mg/liter NaCl feed).

is used to recover wash water from dye ranges and caustic soda from scouring effluents. Dynamic membranes prepared from zirconium hydroxide and polyacrylic acid, for example, are well suited for these applications because they can withstand high temperatures and wide pH ranges, and because performance can be restored by stripping and reforming the membrane in lieu of cleaning.

Many membranes exhibit good rejection toward low concentrations of alcohols, aldehydes, esters, and other organic compounds. However, organic liquid mixtures are as yet seldom separated by reverse osmosis because few membranes developed for desalination exhibit adequate chemical resistance. Moreover, the high osmotic pressures associated with concentrated solutions can drastically reduce the effective driving force and thus the productivity. As an example, fermentation alcohol containing about 8–10% ethanol may be concentrated to only about 60% using present RO technology. As noted earlier, for such applications, the problem of high osmotic pressures can be resolved with another membrane process known as pervaporation (*q.v.*).

A notable shift has occurred over the past decade toward operating RO systems at gradually lower pressures while maintaining the high productivity once associated with high-pressure systems. This is in large part an energy efficiency consideration; lower power consumption will make desalination by RO attractive to a broader range of the global population, for whom the supply of high-quality drinking water will become increasingly critical in the future.

B. Nanofiltration

Since the mid-1980s, ultralow-pressure reverse osmosis—sometimes referred to as “nanofiltration”—systems operating between 5 and 10 bars have gained considerable favor for groundwater softening, organics removal, and even domestic point-of-use water treatment. These systems employ “loose RO” membranes with good rejection toward color substances and organic compounds with molecular weights of several hundred to about 1000 daltons, but only moderately retentive of monovalent salts. These operating characteristics meet the needs for aqueous separations where high productivity and low operating costs are crucial.

Concerns about groundwater contamination and municipal water supply quality have driven much of the growth of various water treatment schemes involving nanofiltration as a stand-alone process or in combination with RO and/or UF in a broad range of water treatment systems delivering precise purity levels and attractive process economics. Other established applications include corn syrup concentration, recycling of water-soluble polymers, effluent treatment for the food and beverage industry, metal

working industry, and organics recovery (e.g., ethylene glycol). Over the past decade, the number of applications and the scale of their implementation continue to grow. So has the range of nanofiltration membranes and systems available commercially.

C. Ultrafiltration

UF is a membrane process useful for separating macromolecules according to differences in molecular size and shape. The fundamentals controlling this process, involving hydrodynamic sieving, have been discussed in the earlier section on mechanisms. The membranes used in UF allow free passage of solvent and solutes with molecular weights below several hundred daltons, while retaining species larger than a characteristic molecular weight cutoff (MWCO). MWCO is a semiquantitative way of specifying the size discrimination characteristics of an ultrafiltration membrane (a common definition being that 90% of the solutes with molecular weights exceeding the MWCO would be rejected by the membrane). Substances that are separated effectively by ultrafiltration include colloids, soluble polymers, and dispersions with molecular weights from a few thousand to about 1 million daltons. In general, species whose molecular weights differ by two orders of magnitude or more may be fractionated by ultrafiltration.

Diafiltration is a variation of ultrafiltration, in which fresh solvent is added to the feed solution to replenish the volume ultrafiltered, and in the process washes small molecules such as salts away from the retained macromolecules. Using appropriate replenishing solutions, diafiltration is a common procedure to perform buffer exchange of proteins. Alternatively, a dilute solution may be first ultrafiltered to concentrate the feed material, then diafiltered to purify the retentate. It is sometimes possible to fractionate a mixture of macromolecules by sequential diafiltration with a series of membranes of progressively lower molecular weight cutoff ratings.

Electrocoat paint recovery in the automotive manufacturing and metal finishing industries is a major UF application. Electrocoating refers to the process of depositing electrophoretic paint from an aqueous dispersion onto immersed, charged metal surfaces. Thin coatings with uniform coverage in recessed areas are obtained. After coating, the metal part is freed of excess paint by rinsing with water. To help the process operate consistently, the paint dispersion is continually purified through an ultrafiltration loop as shown in Fig. 31. Water containing accumulated salts and additives is removed, and the recycled paint is reconstituted with fresh water and solvent and returned to the immersion tank. In this way, UF reduces the cost of wastewater treatment by minimizing water discharge and recovers valuable paint for reuse. An indication of

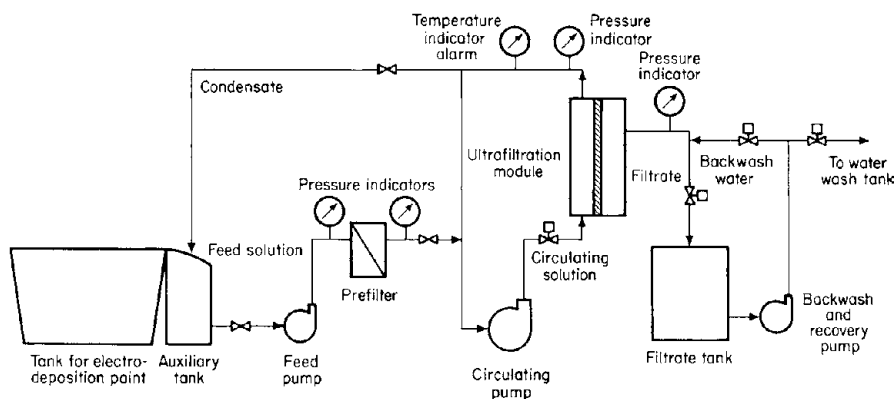


FIGURE 31 An ultrafiltration electrocoat paint recovery system. [Warashina *et al.* (1985, September). *Chemtech*, pp. 558–561.]

the favorable economics is the typical payback period of less than a year for an ultrafiltration installation in modern automotive plants, where adoption of UF technology is practically universal.

The dairy industry was also an early beneficiary of ultrafiltration technology. Major uses include the concentration of whole milk or skim milk for cheesemaking (see Fig. 32); the recovery, fraction, and desalting of whey protein concentrates, and volume reduction of raw milk at the farm to decrease transportation costs. Recently, ultrafiltration has been used to increase the protein content of skim milk. This creates a richer taste that is normally associated with part-skim milk. In other segments of the food processing industry, ultrafiltration is used to concentrate gelatin, clarify fruit juices, removing proteinaceous impurities from wines to improve shelf life, and recover protein from soybean processing and from fishing industry wash water.

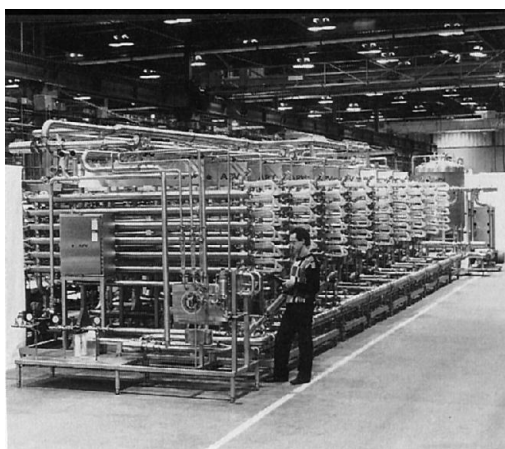


FIGURE 32 A nine-stage ultrafiltration plant concentrating whole milk from 8 to 40% total solids for cheesemaking. (APV Crepeco Inc.)

Another example of using ultrafiltration for wastewater treatment and resource recovery is the separation of oil–water emulsions generated from metal machining, oil field wastes, and enhanced oil recovery effluents. Hydrophilic membranes such as cellulose acetate are preferred because they are effective barriers to oil droplets and are less prone to fouling. The UF permeate readily meets direct discharge standards. The oil-rich stream can be processed to reclaim the oil, or disposed at reduced transportation cost because of its reduced volume.

In the petroleum industry, dewaxing solvents are separated by ultrafiltration from dewaxed oils by chemically resistant membranes made from polysulfone or polyimide. In a related process, pentane is separated from deasphalted heavy oil under conditions intermediate between reverse osmosis and ultrafiltration (ca. 15 bar applied pressure). High-molecular-weight hydrocarbons in the oil form a gel layer on the surface of a polysulfone support membrane. This gel restricts passage of heavier hydrocarbons but not pentane, which is recovered as permeate. To separate other hydrocarbon mixtures that do not contain gel-forming components, polymeric additives would be used as a rejecting barrier substitute.

Textile sizing agents such as polyvinyl alcohol may also be reclaimed from hot process water. Here, both polymeric membranes and inorganic, dynamic membranes are appropriate choices. Systems based on polymeric membranes operate at lower fluxes and require less recirculation pumping, and are somewhat more economical. Plants with treatment capacities as high as 60 m³ per hour are in operation.

Several important UF applications are still in the development stage. For example, metal recovery from plating wastes has been proposed by using a flocculant or a chelating polymer to bind the metal ions, then recovering the polymer complex by ultrafiltration. The metal value may

be reclaimed by smelting, or decomplexed as a concentrated solution, and regenerating the polymer for reuse.

The pulp-and-paper industry is a larger consumer of water: about 70 tons of effluent water is generated for each ton of paper produced from wood pulp. An ultrafiltration system can potentially remove organic materials and reduce biological oxygen demand in the effluent stream, thereby helping compliance with increasingly stringent effluent discharge regulations. A specific opportunity exists in the concentration of black liquor, an alkaline solution laden with lignin and other organics from the kraft pulping process. Black liquor is concentrated at present by flash evaporation and incinerated for its fuel value, but the heat generated only marginally exceeds that required for evaporation. While the ultrafiltration system may improve the energy balance, the membrane materials must be capable of stable operation in the hot and corrosive environment.

Better membrane materials have gradually appeared over the past several years. Ceramic, carbon, and metallic membranes first introduced as microfilters are now commercially available in the ultrafiltration pore size range (ca. 40–1000 Å). They are dominating small but significant markets where their thermal and chemical resistance capabilities are enabling features. For many applications, though, the high cost of inorganic membranes still deters their deployment. Investment in special module housings and membrane geometries discourages replacement even as performance ultimately becomes marginal, as in the case of irreversible fouling.

D. Microfiltration

MF membranes are finely porous, with nominal pore sizes ranging between 0.01 and 5 μm . Some of these membranes are isotropic, i.e., uniformly porous throughout their thicknesses; others have an asymmetric, graded porosity structure. Yet others have more unique morphologies. For example, track-etched membranes are characterized by straight cylindrical pores of uniform diameter; they are made by irradiating thin substrates, then etching away the irradiated paths where the local chemical resistance has been reduced. Biaxial orientation of polymer films or fibers produces microporous membranes with connecting fibrils within each pore. Anodized aluminum membranes with a high density of straight, closely packed uniform pores have also been fabricated successfully.

Separation takes place in microfiltration primarily between solids and liquids, and many established applications are simply extensions of conventional filtration into a lower particle size range. (See Section I.A.) A homogeneous porous membrane used as a conventional depth filter traps particles on its surface and inside the tortuous pores. The membrane can become clogged

rapidly and irreversibly. Pore plugging is reduced with asymmetric microfilters where penetration of particulates below the membrane surface is reduced. Plugging can be further decreased by operating in the crossflow mode. Depending on the application, microfiltration systems may be designed for crossflow or dead-end operation. Fluid management is more flexible in crossflow operation, where high shear conditions can reduce concentration polarization and pore plugging. On the other hand, a higher recovery of the feed fluid is possible with dead-end microfiltration. Dead-end operation is also preferred for processing shear-sensitive feed materials such as certain biomaterials. As with ultrafiltration, the transport properties of the membrane can be strongly affected by concentration polarization, fouling, and interactions between the feed stream and the membrane.

Microfiltration membranes are treated as single-use, disposable items in many clinical, analytical, and laboratory-scale applications where the high value of the product or procedure justifies frequent membrane replacement, and/or the risks associated with reusing contaminated membranes are unacceptable. Membranes used in large-scale industrial MF systems are more often rejuvenated at regular intervals to maximize service life.

The largest microfiltration application is for sterile filtration, or removal of microorganisms, in the pharmaceutical and biotechnology industries. Owing to the high value of the materials being processed, MF is deployed exhaustively and prophylactically, leading to a substantial market size and correspondingly large revenue base. Similarly, MF is used extensively for clarifying fermentation broths as a component of an overall product recovery and purification scheme (see Sections VI and VII).

Food and beverage processing represents an expanding area for process-scale microfiltration. Already in place are clarification systems for wine and beer, sugar, and gelatin, replacing existing practices such as diatomaceous earth filtration. Less attractive economically are miscellaneous waste treatment applications, for which microfiltration is often a sophisticated but expensive alternative.

In semiconductor manufacturing, very-large-scale-integration (VLSI) technology and high-density integrated circuits are made by repeated deposition of extremely fine patterns on silicon wafers. Between process steps, the wafers are cleaned using ultrapure water. The demand for increasing circuit density corresponds directly to increasingly sophisticated water treatment system designs that involve multiple stages of reverse osmosis, ultrafiltration, microfiltration, as well as other nonmembrane technologies. A typical integrated water supply system is illustrated in Fig. 33. Microfiltration of electronics chemicals also represents a large application area within the electronics industry.

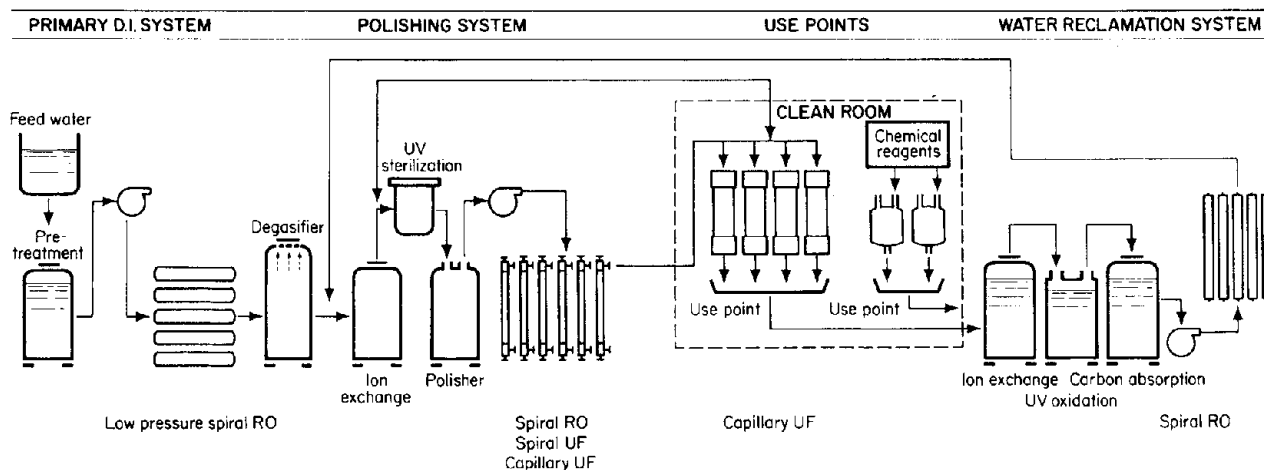


FIGURE 33 An ultrahigh-purity water system for semiconductor manufacture. (Nitto Electric Industrial Co., Ltd.)

E. Membrane Extraction Processes

Membrane extraction encompasses a class of liquid-phase separations where the primary driving force for transport stems from the concentration difference between the feed and extractant liquids rather than a pressure gradient, as is the case with most of the other processes discussed above. A microporous membrane placed between the feed and the extractant liquids functions primarily as a phase separator. The degree of separation achievable is determined by the relative partition coefficients among individual solutes. This operation is known as “membrane solvent extraction.” If a nonporous, permselective membrane is used instead, however, the selectivity of the membrane would be superimposed on the partitioning selectivity; in this case the process may be referred to as “perstraction.” These “process” concepts are illustrated in Fig. 34.

1. Membrane Solvent Extraction/Membrane Contactors

Conventional liquid–liquid extraction is an established unit operation for transferring one or more solutes from a solution into a second, immiscible liquid. It is widely used for separating ionic and nonionic species, for example, on the basis of their preferential partitioning between an aqueous phase and a nonaqueous phase, respectively. Industrial liquid–liquid extraction equipment generally consists of a mixer, where the feed solution and the extractant liquid are intimately mixed via agitation, and a settler where the equilibrated phases are separated for further processing.

Phase separation may or may not occur spontaneously after mixing. If surface-active species are present, for example, the mixed phases may remain dispersed for long

periods of time. In extreme cases an emulsion may form that is indefinitely stable. Whenever phase separation is incomplete, there is entrainment loss of one solution in another. In addition to low overall separation efficiency, valuable products or extracting agents are lost.

Using a solid, microporous membrane to define a stationary phase boundary during extraction may alleviate this problem. The feed solution and the extractant flow

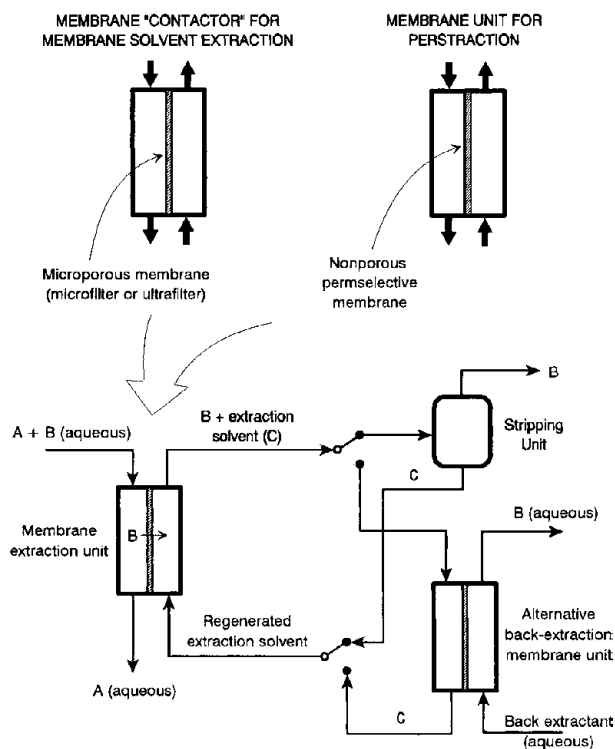


FIGURE 34 Membrane extraction processes: (a) membrane solvent extraction and (b) perstraction.

on opposite sides of the membrane and contact through its micropores, where mass transfer takes place. Dispersion/emulsification problems are avoided since the bulk solutions do not mix. By using membranes with high packing densities, e.g., hollow fibers, a large phase contact area can be obtained per unit volume. The most commonly prescribed membranes for this purpose are polyolefin hollow fibers. There are many liquid–liquid extraction applications for which membrane solvent extraction is a viable alternative or an enabling technology. Thus far, however, there are few known examples of commercial-scale membrane solvent extraction, due mostly to the relatively high cost of membrane systems compared to mixer/settler equipment. A second reason is the lack of suitable membranes that are solvent-resistant and have pores small enough not to allow breakthrough of one phase into the other under modest pressure imbalances, or slow but nonnegligible emulsification. A viable alternative is polyacrylonitrile hollow-fiber membranes with pore sizes normally associated with ultrafiltration membranes. With their good solvent resistance and a reduced tendency for phase breakthrough, these membranes hold the promise as a generic membrane solvent extraction tool.

2. Perstraction

Perstraction is a process analogous to pervaporation, except that a liquid extractant is used instead of a partial vacuum or sweep gas to carry the permeate away from the permselective membrane. The liquid extractant is regenerated by passage through a stripping device. In principle, perstraction offers the potential of higher selectivity than those achievable by liquid–liquid extraction or membrane solvent extraction. To maximize the effectiveness of this approach, the membrane should be chosen such that its permselectivity is complementary or additive with the equilibrium partitioning properties of the feed solution/extractant pair. In practice, with the exception of ethanol–water separation, the promise of additive selectivity is not well exploited to date because of the considerable development effort required to optimize a given separation. Successful applications will likely be limited to separations of high-value products for which the development of a unique permselective membrane for a single purpose can be justified.

F. Liquid Membranes

Permeation through liquids is orders of magnitude faster than that through solid polymers of comparable thickness. This rate advantage is exploited for some separations by using an immiscible liquid film as the membrane to mediate the transport of selected substances. Two somewhat different separation technologies have evolved based on

this principle: emulsified liquid membranes, where discrete encapsulated droplets serve as selective reservoirs for certain species in the surrounding solution, and immobilized liquid membranes, where a microporous solid support holds the liquid as a continuous barrier between the feed and permeate streams. Both are intimately related to conventional solvent extraction in the selection of extractants and the physical chemistry of the process. As further refinements of this configuration, selective carriers may be incorporated into the immobilized liquid to enhance extraction selectivity. Processes variously referred to as “facilitated transport” and “coupled transport” are examples of this approach.

1. Emulsified Liquid Membranes

A liquid membrane can be prepared by emulsifying an aqueous solution in an organic liquid, then adding the emulsion to another aqueous solution. In this way, the organic liquid segregates the solutions but allows selective diffusion of solutes across it. Similarly, oil/water/oil type emulsions can be formed in which the liquid membrane is the water encapsulation layer. Very high rates of mass transfer can be achieved because of the large effective membrane surface area represented by the emulsion droplets.

Separation in liquid membranes can take place in several ways, as shown in Fig. 35. The simplest mechanism (a) is selective partition of solutes from the first aqueous phase into the encapsulating organic liquid, followed by selective desorption into the second aqueous phase. Dissolved hydrocarbons have been separated using this approach. However, the extraction capacity of each membrane-encapsulated droplet is limited by its size because the thermodynamic activity inside the droplet cannot exceed that in the feed. Backdiffusion can be prevented by chemically converting the extracted solute (b) so as to maintain the driving force for diffusion of unconverted solute. To extract phenol from wastewater, for example, a liquid membrane prepared by encapsulating sodium hydroxide solution in a hydrocarbon liquid is used. Phenol reaching the sodium hydroxide is converted into phenolate ions, which is virtually insoluble in hydrocarbons and cannot backdiffuse into the feed solution. A similar approach can be used in general to recover organic acids that partition readily into hydrocarbons as neutral molecules and accumulate in dissociated form in the encapsulated liquid. Even more complex reaction strategies may be implemented as shown in mechanism (c). At this time, however, there are relatively few liquid membrane extraction systems in commercial use.

The equipment used for emulsified liquid membrane extraction, shown in Fig. 36 for a wastewater treatment

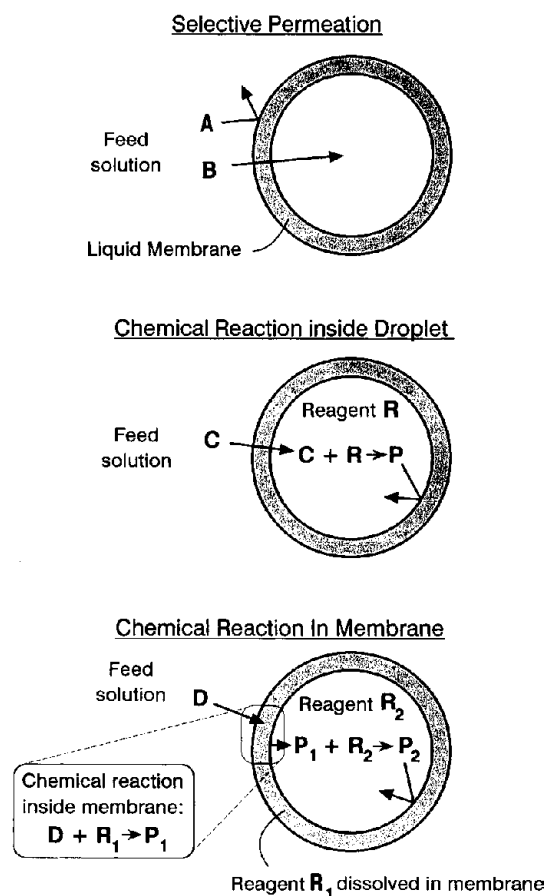


FIGURE 35 Emulsified liquid membrane separation mechanisms: (A) selective permeation; (B) chemical reaction inside emulsion droplet; and (C) chemical conversion in liquid membrane and further conversion inside droplet. Both (B) and (C) provide quasi-infinite sink conditions for extraction from the feed solution.

system, bears much resemblance to conventional liquid-liquid extraction systems. The liquid membrane formed is mixed with the feed water to allow extraction to occur, then decanted off after the liquid membrane is saturated with

the target substances. The emulsion is coalesced chemically or electrostatically to release the encapsulated liquid and to recycle the liquid membrane constituents.

2. Immobilized Liquid Membranes

An immobilized liquid membrane is formed by impregnating a microporous support with an extractant liquid. The liquid is held in place by capillarity and assumes the flat-sheet or hollow-fiber geometry of the host membrane. Immobilized liquid membranes can be used for virtually all the liquid-phase separations achievable with emulsified liquid membranes, but offer several important benefits. There should be no entrainment loss because no mixing occurs. Also, extraction and stripping of target species occur simultaneously on the upstream and downstream surfaces of an immobilized liquid membrane, respectively. The size of the receiving phase can thus be virtually unlimited by continually regenerating and recycling the stripping solution. Hollow-fiber devices may be used to favor a high packing density of contact area between the immiscible phases. Finally, because it is supported in a solid matrix, an immobilized liquid membrane is applicable to the separation of gases and vapors.

3. Facilitated Transport and Coupled Transport

It is possible to achieve very high selectivities by incorporating complexing agents or carriers in immobilized liquid membranes. These agents may be liquid ion exchangers or chelating polymers; they form reversible complexes with the target species on the feed side of the membrane and release those species by dissociation on the downstream side. As the overall selectivity of this process depends on the specificity of chemical recognition—sometimes at low concentration and often in the presence of interfering species—much effort has been focused on developing sophisticated complexing agents such as macrocyclic

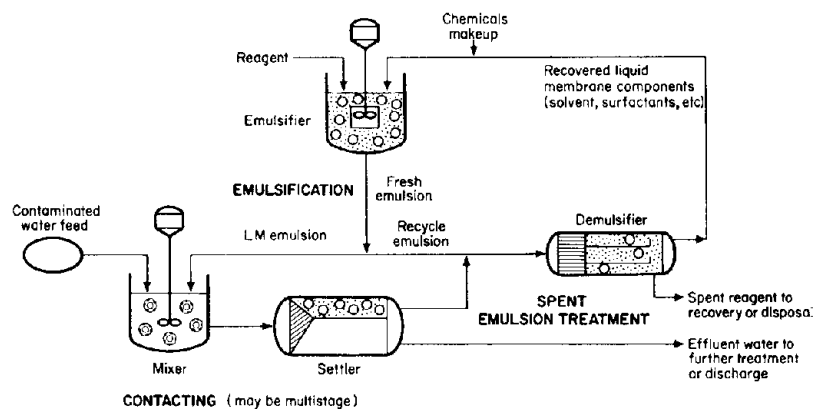


FIGURE 36 An emulsified liquid membrane wastewater treatment process.

compounds (e.g., cyclodextrins) with well-defined cavity sizes or those carrying coordinating functional groups.

A special case of facilitated transport involves the use of organic-soluble liquid ion exchangers to recover metal ions from dilute solutions. Often referred to as coupled transport, this process operates by driving the transport of the metal complex with the flow of a second species (most often protons in the form of a pH gradient) in the opposite direction. As depicted in Fig. 37, coupled transport can operate by two mechanisms: (1) *cotransport*, where metal-containing anions permeate in the same direction as the protons, and (2) *countertransport*, where metal cations and protons (or analogously, metal-containing anions and another anion supplied from the stripping solution) permeate in opposite directions. In all cases, the pH of the external solutions is adjusted to provide favorable conditions for the complexation and decomplexation reactions at the solution–membrane interfaces, and to maintain the pH gradient as driving force. Very clean separations are possible in extracting metal ions from dilute solutions, or in separating two or more metal ions with different complexation characteristics. Practical applications in the plating and metal-finishing industry, wastewater treatment, and hydrometallurgical extraction of ores have been contemplated. Until recently, however, commercialization of this technology seems to be hampered by the fluctuating prices of metals such as chromium and copper, or by uncertainties in the commodity value of uranium (Ho, 2000).

Most liquid membranes are less stable than their polymeric counterparts. Although the thin liquid film in the membrane corresponds to a short diffusion path and hence a high mass transfer rate, small amounts of the immobilized liquid can be displaced under pressure. Also, the immobilized liquid may slowly dissolve in the external phases, eventually leading to discontinuities in the liquid

film. Limited lifetime is perhaps the most important liability against practical application of this technology. To address this problem, experimental membranes containing high concentrations of complexation sites in a solid polymeric matrix have been developed. Above a certain critical carrier density, the transport of the complexed species can take place by site-to-site jumps—a “bucket brigade” effect. Because the complexation sites are an integral part of the polymer, there is little loss of efficiency so long as the host polymer remains stable.

G. Industrial Dialysis, Donnan Dialysis, and Electrodialysis

1. Industrial Dialysis

Dialysis operates by the diffusion of selected solutes across a nonporous membrane from high to low concentration. An early industrial application of dialysis was caustic soda recovery from rayon manufacturing. It had been a viable process because inexpensive but alkali-resistant cellulose membranes were available that were capable of removing polymeric impurities from the caustic. Gradually however, dialysis is being replaced by dynamic membrane technology for caustic soda recovery because of the latter’s much higher productivity.

Dialysis continues to meet certain specialized applications, particularly those in biotechnology and the life sciences. Delicate substances can be separated without damage because dialysis is typically performed under mild conditions: ambient temperature, no appreciable transmembrane pressure drop, and low-shear flow. While slow compared with pressure-driven processes, dialysis discriminates small molecules from large ones reliably because the absence of a pressure gradient across the membrane prevents convective flow through defects in the membrane. This advantage is significant for two

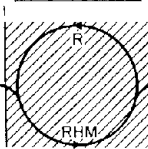
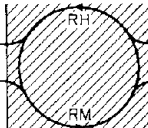
Process	Mechanism			Species transported
	Feed solution	Liquid membrane	Product solution	
Cotransport	$H^+ + M^-$ Low pH		$H^+ + M^-$ High pH	M^- : $Cd(CN)_4^{2-}$, CrO_4^{2-} , $Cr_2O_7^{2-}$, $Cu(CN)_2^-$, $MnCl_3^-$, $UO_2(SO_4)_2^-$, $UO_2(SO_4)_3^{4-}$
Countertransport	H^+ High pH M^+		H^+ Low pH M^+	M^+ : Al^{3+} , Cd^{2+} , Co^{2+} , Cr^{3+} , Cu^{2+} , Hg^{2+} , MoO_4^{2-} , Ni^{2+} , Pb^{2+} , UO_2^{2+} , Zn^{2+}

FIGURE 37 Mechanisms of carrier-facilitated immobilized liquid membrane extraction, also referred to as coupled transport. The species, R, refers to the carrier component responsible for complexation.

reasons. The first relates to critical applications—e.g., medical/immunological separations and salt removal from solutions of genetically engineered proteins—where leakage of undesirable species from the feed stream into the permeate cannot be tolerated. (Also see Sections VI and VII.) The second aspect is the absence of concentration polarization arising from convective flow through an ultrafilter, for example, and the consequent accumulation of rejected species in the boundary layer.

2. Donnan Dialysis

Ion exchange membranes contain high concentrations of fixed charges. They are permeable to ions of opposite charge (counterions) but repel ions of the same charge (coions). Protons are the only exception; they can permeate freely through hydration passages in an anion exchange membrane. The functions of anion- and cation-exchange membranes are illustrated in Fig. 38.

Donnan dialysis functions through the interaction between ions and ion-exchange membranes in the absence of an externally applied electrical field. When an ion exchange membrane separates two electrolyte solutions, and a second electrolyte with the same counterion but a nonpermeating coion is added to one side of the membrane, counterions migrate across the membrane until the charge separation stops further flow and electroneutrality is established on both sides of the membrane. This phenomenon is known as Donnan equilibrium. Donnan dialysis refers to the process of separating ionic components in a feed stream according to their tendency to migrate across ion-exchange membranes to achieve equilibrium.

The example shown in Fig. 39 illustrates the treatment of an aluminum anodizing bath waste stream by Donnan dialysis. Sulfate ions and protons freely permeate from a feed stream of aluminum sulfate and sulfuric acid across

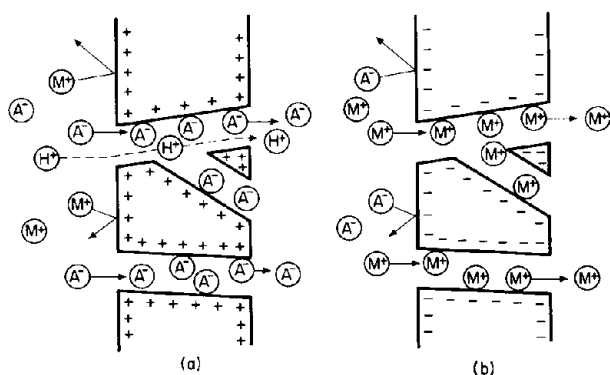


FIGURE 38 Selective diffusion across ion-exchange membranes. (a) Anion exchange, and (b) cation exchange. Metal cations are designated by M^+ , anion A^- , proton H^+ , and the fixed charges in the membrane by $+$ and $-$.

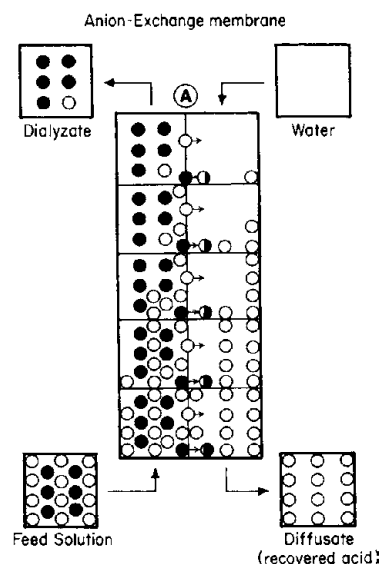


FIGURE 39 Donnan dialysis application to the separation of sulfuric acid from aluminum sulfate. $Al_2(SO_4)_3$ designated by $\circ$ and H_2SO_4 by $\bullet$ (HPD, Inc.).

the ion exchange membrane into a water stream, forming sulfuric acid. Aluminum cations are rejected by the fixed positive charges on the membrane and exit as a less acidic aluminum sulfate stream for recovery or disposal. Similar applications include the recovery of sulfuric acid from nickel sulfate steel pickling waste, and the recovery of nitric and hydrofluoric acids produced during stainless steel etching. Donnan dialysis is effective because high concentration gradients yield concentrated products, and because direct input of electrical energy is not required to achieve separation.

3. Electrodialysis

Although the development of electrodialysis desalination technology predated that of reverse osmosis (*q.v.*), at present both processes compete favorably with distillation for potable water production. In electrodialysis, salts are removed from a feed solution by using an electric current (DC) to transport ions across anion-exchange and cation-exchange membrane pairs. By restricting the migration of ions to no more than one adjacent solution compartment, as shown in Fig. 40(a), alternate streams become enriched and depleted of electrolytes. Electrodialysis operates most economically when the feed water contains less than 0.5% TDS, but medium-salinity seawaters (up to about 1.2% TDS) can also be desalted. Product water containing less than 0.01% TDS can be obtained. The capability of electrodialysis to remove salts from neutral solutes is also exploited in other applications, e.g., desalting proteins.

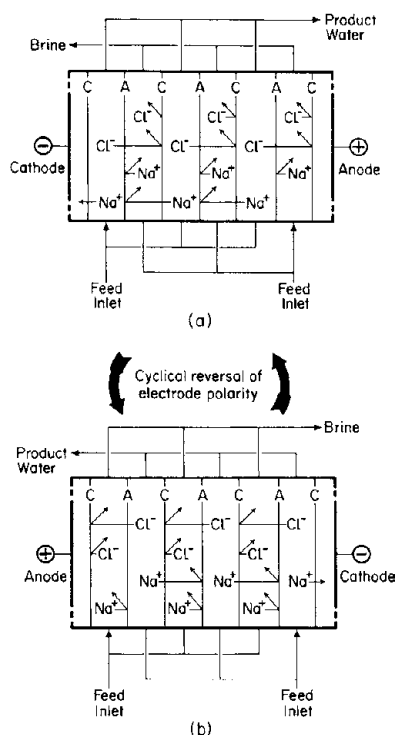


FIGURE 40 (a) Electrodialysis and (b) electrodialysis reversal (EDR). Cation exchange membrane indicated by C and anion exchange membrane by A (Ionics Inc.)

In electrodialysis, membrane fouling occurs after relatively short periods of operation because scale-forming ions migrate unidirectionally to the membrane surface. The result is gradually increasing electrical resistance and reducing desalting efficiency. Pretreating the feed water with chemicals delays fouling but does not prevent it. Electrodialysis Reversal (EDR) was a process improvement introduced in the early 1970s that overcame the fouling problem by reversing the polarity of the DC field at 15- to 20-min intervals, as shown in Fig. 40(b), and purging the removed scale and foulants from the stack. Today EDR is a proven process for brackish water desalination: one of the largest commercial installations in Florida, United States, produces 12 million gallons of drinking water per day from 0.13% TDS feedwater at 85% recovery.

H. Electrochemical Synthesis and Bipolar Membrane Technology

The unique capability of ion-exchange membranes to separate chemical species according to ionic charge makes it possible to conduct various electrochemical synthesis reactions otherwise difficult to perform. A number of such synthetic mechanisms are shown in Fig. 41. Although each may be applied individually, a recent trend has emerged toward assembling several electrochemical

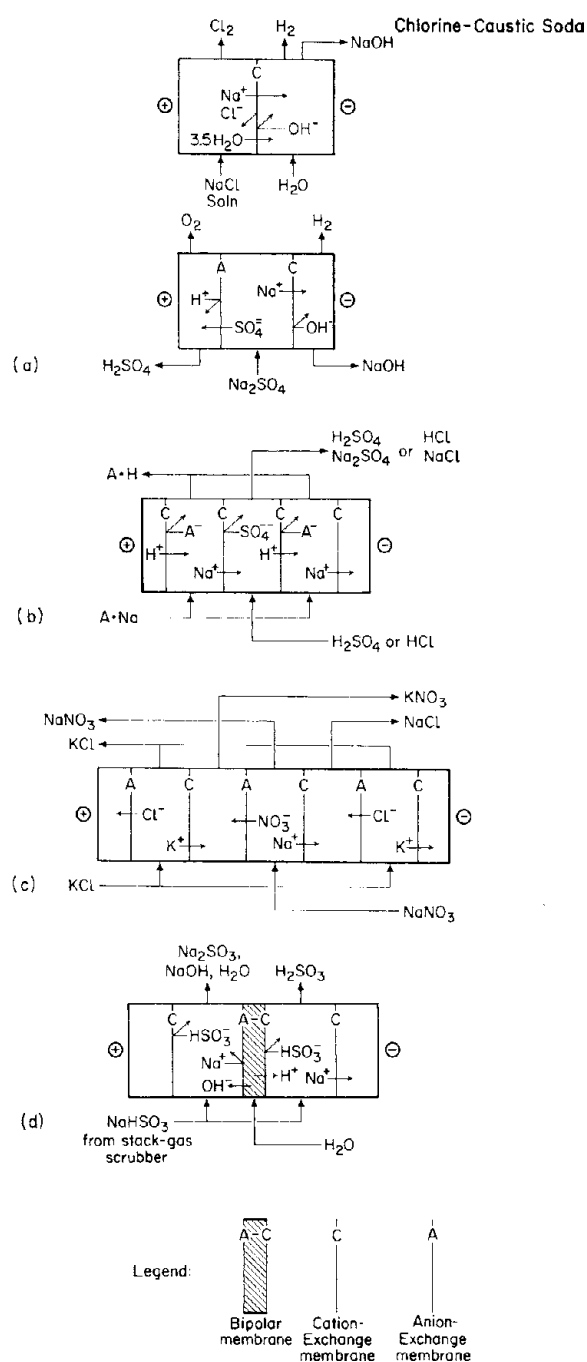


FIGURE 41 Membrane-based electrochemical syntheses: (a) electrolysis; (b) substitution; (c) double decomposition; and (d) bipolar membrane synthesis.

reaction schemes into innovative waste treatment and resource reclamation systems.

1. Electrolysis

The largest scale synthesis based on electrolysis is the chlor-alkali process. Sodium ions in a salt brine migrate

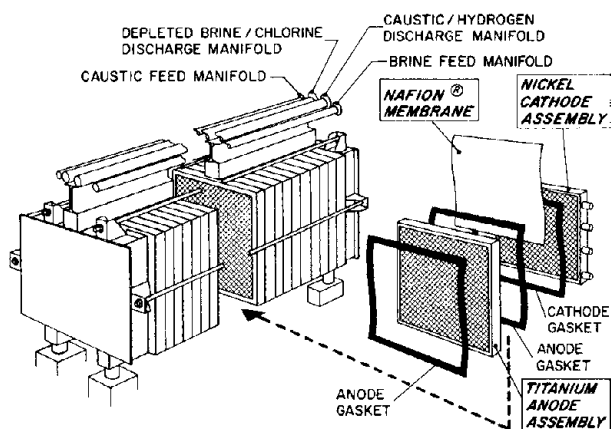


FIGURE 42 Construction of a chlor-alkali membrane unit for electrolysis of brine. (Du Pont Company.)

across a cation-exchange membrane to form caustic soda, and chloride ions react at the anode to form chlorine gas. The construction of a commercial chlor-alkali membrane-cell assembly is shown in Fig. 42.

The use of polyperfluorosulfonic acid membranes as the cell separator was first demonstrated about three decades ago. Yet it was not until the mid-1980s when the economic advantages of membrane cells over the traditional mercury- and diaphragm-cell technology were fully demonstrated—consequent to better membrane performance, higher caustic product concentrations, and lower power consumption. Retrofitting chlor-alkali facilities with membrane cells accounted for much of the growth and sustenance of this industry over the past two decades.

By forming an electrolytic cell with both an anion-exchange membrane and a cation-exchange membrane, acid and alkali can be generated simultaneously. The method applies to inorganic salts (as illustrated) and organic salts (e.g., sodium citrate converted to citric acid and sodium hydroxide).

2. Substitution Reactions

In substitution reactions, solutions of a salt and an acid with the same anion are fed through alternate compartments of an array of cation-exchange membranes. The dissociated metal ions from the salt are removed and replaced by protons to generate the free acid. For example, amino acids are produced from their sodium salts in this way. Compared with conventional neutralization and recovery techniques, the membrane-mediated process is considerably simpler and gives a higher yield of the purified product.

3. Double Decomposition

Double decomposition is similar in concept to the substitution reaction, except that both anion-exchange and cation-exchange membranes are employed. Simultaneous interchange of anion–cation pairing takes place to form products that would otherwise require multistep procedures to prepare and purify. Pure materials can be produced from crude raw materials by means of double decomposition, and reactions otherwise impractical by conventional reaction methods can be performed. An example application is the reaction between potassium chloride and sodium nitrate to produce potassium nitrate and sodium chloride.

4. Bipolar Membrane Syntheses

A bipolar membrane consists of a cation-exchange layer and an anion-exchange layer, separated by a thin water-filled space. Placing this membrane between cation-exchange membranes and electrodes in the orientation shown in Fig. 41(d) forms a special electrochemical cell. When direct current is passed through the cell, water between the two layers of the bipolar membrane electrolyzes to release protons and hydroxyl ions into adjoining compartments, where they participate in substitution reactions. Bipolar membrane technology may be considered a second-generation electrochemical synthesis because of its versatility: different arrangements of bipolar membranes together with cation- or anion-selective membranes can separate a salt into its constituent acid and base, or produce purified acid or base streams. Several of these schemes are shown in Table IX.

The schematic shown in Fig. 43(a) is a commercial example of this technology. Stack gas is scrubbed with an alkaline solution of sodium hydroxide, sodium sulfite, and sodium sulfate. The sodium sulfite reacts with SO_2 in the stack gas to form sodium bisulfite. This salt solution is processed in a bipolar membrane unit [Type (I) shown in Table IX] to generate an alkaline solution and an acidic solution. The alkaline solution contains regenerated caustic soda and sodium sulfite, and can be recycled to the scrubber, while the sulfurous acid can be further processed to sulfur or sulfuric acid for sale.

Bipolar membrane synthesis also holds promise for regenerating spent pickling liquors in stainless steel manufacture. As shown in Fig. 43(b), waste acid laden with metal ions can be continuously neutralized, filtered to remove the precipitated metal oxide, and the clarified salt solution split into its acid and base components in a bipolar membrane unit [Type (IV) shown in Table IX]. As much as 95% of the hydrofluoric and nitric acid used are returned to the pickling bath, thereby solving a waste

TABLE IX Bipolar Membrane Electrochemical Synthesis Schemes

Scheme	Cell/membrane arrangement	Features
(I) Two-compartment cation cell		Converts concentrated weak acid salt solution to pure base solution plus mixed acid/salt stream
(II) Two-compartment anion cell		Converts weak base salt solution to salt/base mixture plus pure acid
(III) Multichamber cation cell		Higher acid concentration in the salt/acid stream than that for Scheme I
(IV) Multichamber anion cell		Higher base concentration in salt/base product than in Scheme II

disposal problem while minimizing the consumption of fresh acid.

I. Catalytic Membrane Reactors for Chemical Processing

There are a number of advantages of using membrane systems to conduct chemical reactions or syntheses. A single device could in principle integrate reaction, concentration, and separation functions. Segregating reactants from products would also enhance thermodynamically limited or product-inhibited reactions. Initially, the lack of membrane materials sufficiently resistant to high temperatures and chemical attack precluded the realization of mem-

brane reactor concepts in much of the petrochemical and chemical process industries.

During the 1990s, advanced fabrication methods were developed to convert ceramic and other inorganic materials into membrane structures with a range of pore size and form factors. As membranes with Ångstrom-size pores are developed, reaction and separation in the gas phase may be accomplished. Complementary technology has also been developed for catalyst loading into membrane matrices. Much of this progress has come from companies with a basic position in inorganic membrane and catalysis technologies.

The most interesting candidate reactions for chemical membrane reactors will be those currently compromised

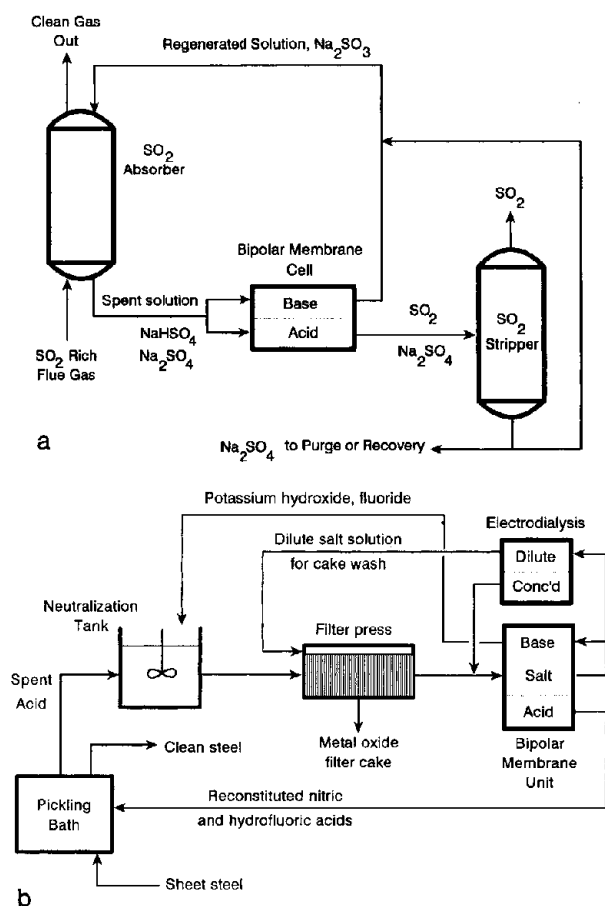


FIGURE 43 Bipolar membrane processes for (a) SO₂ removal from stack gases; and (b) stainless steel pickling bath waste acid regeneration.

by harsh conditions, unfavorable kinetics, catalyst poisoning, ineffective removal of inhibitory products or intermediates, and/or troublesome product recovery. Some process concepts are shown in Table X.

VI. BIOTECHNOLOGY AND LIFE SCIENCES

Over the past quarter century, biotechnology has fundamentally transformed the life sciences from operating within the confines of nature to a point where manipulation of the structure and behavior of life forms have become both routine and commercially successful. Advances in recombinant DNA technology, for example, have enabled production of highly effective vaccines, antibodies, growth hormones, and other biopharmaceuticals. Even more recently, the Human Genome Project has begun to yield important—if not yet complete—information about the genetic origins of diseases, enabling extremely focused development of therapeutic countermeasures. The

phenomenal growth of this industry has also brought with it new challenges in the areas of separation and purification. Membrane technology offers a number of existing solutions and promises new ones. In this section, the use of membranes as tools in various stages of life science research up to large-scale production of biopharmaceuticals will be examined.

Certain special requirements apply to membrane systems used in the life sciences. Proteins, cells, and their constituents retain their biological functions within a relatively narrow range of environmental conditions. Many are sensitive to provocation or damage when those conditions change, or even upon contact with surfaces recognized to be foreign. For these reasons, materials used to separate or purify biological materials must be “biocompatible” to various extents, and process conditions established to avoid irreversible changes in the desired product. An entire area of study has emerged focused on the development and optimization of biomaterials for different purposes, including those used to prepare or modify synthetic membranes.

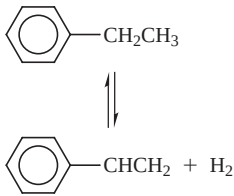
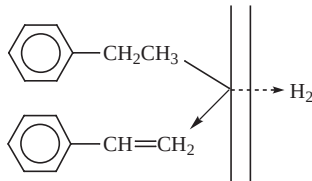
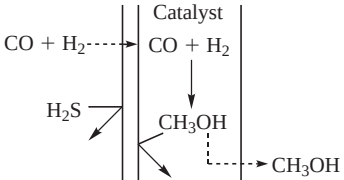
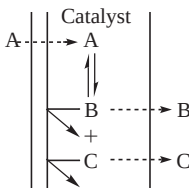
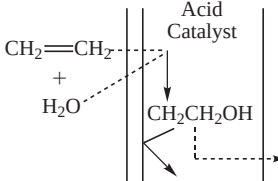
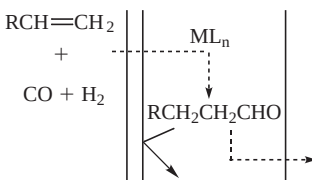
A. Applications in Discovery and Research

Living systems are enormously complex in their composition and function. Understanding individual interactions frequently begins with resolving, identifying, and quantifying key components of interest. Many laboratory procedures consist of steps aimed at recovering a single biological or biochemical entity in high purity. Microfiltration is routinely used for separating particulates, cells, and cell fragments from soluble proteins. Ultrafiltration has also become the preferred procedure for desalting proteins, nucleic acids, or peptides. Proteins with different molecular weights may also be fractionated. A common technique for resolving protein mixtures is electrophoresis (*q.v.*), in which a sheet of hydrogel carrying the protein mixture is subject to an electrical field to cause differential migration according to the charge characteristics of each component. At the conclusion of the electrophoretic separation, a microporous membrane is often applied (blotted) onto the hydrogel to transfer the pattern of resolved proteins, nucleic acids, and their fragments to a stronger substrate to facilitate further analysis or handling.

Laboratory membrane applications usually involve small samples. Consistency and resolution of the separation is as important as productivity of the membrane. As is typical in analytical work, membranes or membrane devices are usually used once and discarded. Strict compliance to sterility and validation requirements is expected.

Various technological advances have contributed to the success to date of genomics, the deciphering and systematic investigation of genetic information embedded

TABLE X Catalytic Membrane Reactor Concepts for Chemical Synthesis

Reaction type	Example	Conditions	Limitations	Membrane concept
Dehydrogenation	Ethyl benzene to styrene 	500–600°C; Fe-Cr-K oxide catalyst	Equilibrium limited conversion Costly separations Endothermic	
Hydrogenation	Carbon monoxide to methanol $\text{CO} + 2\text{H}_2 \rightleftharpoons \text{CH}_3\text{OH}$	250°C; 50–100 bar; Cu-Al-Zn oxide catalyst	Equilibrium limited conversion (12–15%) Catalyst poisoning (by S, Cl) Exothermic	
Olefin metathesis	$\text{RCH}=\text{CH}(\text{CH}_2)_n\text{COOR} \rightleftharpoons \text{RCH}=\text{CHR} + \text{CH}(\text{CH}_2)_n\text{COOR}'$	25–400°C; W, Re (homogeneous, heterogeneous)	Equilibrium mixtures require product-feed separations Catalyst poisoned by water	
Hydration	Ethylene to ethanol $\text{CH}_2=\text{CH}_2 + \text{H}_2\text{O} \rightleftharpoons \text{CH}_3\text{CH}_2\text{OH}$	300°C; 70 bar; H_2PO_4 catalyst	Equilibrium limited conversion (<5%) Large recycle	
Hydroformylation	$\text{RCH}=\text{CH}_2 + \text{CO} + \text{H}_2 \rightarrow \text{RCH}_2\text{CH}_2\text{CHO}$	100–200°C; 200–450 bar; Co, Ru, Rh homogeneous complex catalysts	Catalyst recovery Product separation	

in genomes, and proteomics, the comprehensive study of proteins. Some examples are (1) increased automation, through the use of robotic systems and powerful computers, to conduct massive screening of candidates to identify those showing an expected response; (2) miniaturization of “classical” membrane processes; and (3) improved analytical sensitivity and specificity such as those associated with PCR (polymerase chain reaction) or special mass spectrometry techniques. With respect to membrane systems, 96 individual assays may be performed systematically in as many cavities built into a multiwell form factor (Fig. 6). The bottom of each cavity may be sealed with a microfiltration membrane, for example, to retain cells but permit nutrients or reagents to perfuse through, thereby

creating a miniature bioreactor chamber. If an ultrafiltration membrane is used instead, proteins expressed by the cells could be retained and subsequently desalted and purified by successive buffer exchange. Biochemical reactions may be conducted *in situ* and monitored instrumentally by fluorescence, luminescence, or absorbance measurements. In response to the ever-increasing pace of drug discovery driven by pharmacogenomics, new designs of membrane microplates containing 384 or even 1536 wells have been introduced to support rapid parallel processing. There is little doubt that traditional laboratory membrane operations will be adapted to this large-scale integration strategy, analogous to the evolution of integrated circuits in the microelectronics industry.

B. Process Bioseparation

At the end of successful discovery and clinical trial phases of a biopharmaceutical development program, the goal is production on a scale large enough to be commercially attractive. Through a series of scale-up operations, a laboratory-scale procedure is transformed to process-scale equipment and protocols. Membrane processes have been shown to be readily scalable—i.e., the performance of large systems may be accurately predictable from the behavior of small systems—and economically competitive with other means of separation such as chromatography. Today, membrane processes have become a broadly adopted unit operation in the biotechnology industry in which as much as 90% of the overall manufacturing cost may be attributed to separation and purification operations. A number of membrane processes that are suitable alternatives to conventional processing techniques are listed in Table XI. A recent reference focusing on filtration in biopharmaceutical production covers the use of membranes and depth filters, including detailed monographs on various sectors of this regulated industry (Meltzer and Jornitz, 1998).

1. Separation by Size or Charge Discrimination

Whenever a large difference in size exists among species to be separated, the task of selecting membrane processes and materials is relatively simple. Often the choice is governed by the characteristic size of the species to be retained. For example, microfilters and ultrafilters are generally used to retain and purify particulates and soluble macromolecules, respectively, as discussed in previous sections. Experience over several decades of laboratory- and process-scale practice has led to some rules-of-thumb. For example, hydrophobic membranes are prone to interaction with various proteinaceous materials. Nonspecific binding of solutes reduces the likelihood of a clean separation; it also increases the risk of fouling and consequent loss of flux. Hydrophilic membranes are therefore preferred for most bioseparations. Similarly, membranes bearing functional groups undergo acid–base equilibrium in response to the pH and ionic strength of its aqueous environment. The resultant charge on the membrane surface interacts with the net charge on solutes ranging from amino acids to proteins whose electrochemical signatures are determined by their isoelectric points. A widely accepted

TABLE XI Conventional Bioseparation Technologies and their Membrane-Based Alternatives

Unit operation	Conventional bioseparation process	Membrane process	Membrane process feature
Cell harvesting	Centrifugation	Crossflow microfiltration (MF) Ultrafiltration (UF)	Separation does not rely on density differences High cell densities obtainable
Whole-broth clarification	Rotary vacuum drum filtration Centrifugation	Crossflow MF, UF	Maintains high throughput Does not require filter aids
Protein concentration and purification	Electrolyte/solvent precipitation Affinity column chromatography	UF Diafiltration Membrane-modulated precipitation Membrane-based affinity separation	Recovery and purification performance controllable via membrane selection Some fractionation possible Much higher throughput Smaller investment in expensive ligands Reduced hold-up volume; thus lower product loss risks Economical scale-up
Desalting	Ion-exchange column chromatography Size-exclusion (gel-permeation) chromatography	Membrane-based ion exchange chromatography Electrodialysis Dialysis UF	High volumetric efficiency Higher throughput for given size of equipment Economical even at large scale
Acid/base recovery	Chemical treatment Ion exchange	Bipolar membrane synthesis	Requires no chemical additive; products may be directly recycled
Microsolute concentration	Vacuum evaporation	Reverse osmosis Pervaporation Membrane distillation	Reduced loss of volatile products Less damage to heat-sensitive substances
Solvent extraction	Podbielniak extraction	Membrane solvent extraction Coupled transport	Minimum emulsification and associated entrainment loss Enhanced selectivity and concentration

notion is that at least one to two orders of magnitude difference in molecular weight between the retained species and the permeated species is required to achieve “clean” separation, as in protein solutions containing electrolytes and small organic molecules.

Clearance of viruses from biopharmaceuticals, blood components, and plasma derivatives is essential for safeguarding against transmission of pathogenic agents. Effective validation of inactivation or removal of viruses is also a regulatory requirement. Membrane filtration is used routinely, both for preventing entry of viruses into bioprocesses as well as clearing them prior to final packaging of the product (Levy, Phillips, and Lutz, 1998; Aranha, 2001).

Virus particles range in size from about 20 to 300 nm in diameter. Since they are larger in size than most proteins, viruses may be segregated by size discrimination using membranes similar to those used for ultrafiltration or nanofiltration. Depending on the concentration of the protein, the virus targets and their sizes, dead-end filtration (also referred to as “normal-flow filtration” or “direct-flow filtration”) or crossflow (tangential-flow) filtration may be more effective. Commercial membranes designed for virus removal may be isotropic or asymmetric; they are used primarily for normal-flow and tangential-flow configurations, respectively. Although they exhibit a range of characteristic pore structures and sizes, virus filtration membranes are usually rated by their ability to reduce the titer of given viruses under given conditions by a log reduction value, $LRV = \log(C_i/C_p)$, where C_i and C_p are the virus concentrations within the feed and permeate streams, respectively. Regulatory guidelines usually recommend a cumulative LRV of 12, or a twelve-log reduction in virus titer, in most protein purification processes. This is usually accomplished by a combination of chromatographic and membrane processes.

2. Concentration Polarization and Hydrodynamic Countermeasures

Other important considerations in process bioseparations are fluid management and membrane rejuvenation methods. Crossflow, or flow tangential to the membrane surface, induces shear at the membrane surface and helps reduce concentration polarization. This flow pattern also creates lift forces that counteract the deposition of particulate matter on the membrane resulting from permeation flow normal to the membrane surface. (See Section I.A.)

An effective method of controlling concentration polarization and sustaining productivity involves inducing turbulent vortices on the membrane surface to counteract the forces of solute or particle deposition. The rotating

membrane separator is an example of this approach. It consists of a membrane mounted on a porous rotor inside a concentric cylinder. In operation, a suspension is introduced into the annular space between the membrane and the outer cylinder while the rotor is driven at high angular velocity. Taylor vortices formed inside the annular space help prevent the suspended solids from adhering to the membrane. The solid components thus collected in the annular space exit in a concentrated stream while the liquid portion passes through the membrane and the core of the rotor. In principle this design can be applied to various separations where concentration polarization is a serious concern, such as cell cultures or fermentation broths (Belfort, Davis, and Zydney, 1994). In practice, the most successful application is in plasmapheresis (See Section VII.B.)

3. High-Performance Tangential Flow Filtration

In biotechnology, one often encounters mixtures of proteins whose molecular weights differ by less than an order of magnitude. Fractionating such mixtures had not been considered feasible by traditional ultrafiltration methods. However, by the mid-1990s a strategy was developed that combines the effects of size discrimination, charge interactions, management of hydrodynamics, and module design to yield exceptional selectivity in separating mixtures of biomolecules whose molecular weights differ by as small a factor as two or even less. Referred to as high-performance tangential-flow filtration (HPTFF), this technique begins with a detailed profiling of the electrochemical properties of each protein to be separated, and then selecting an operating pH to maximize the difference in the net charge—hence accentuating the difference in the coiled or extended conformation—of the proteins. A membrane is chosen whose overall pore size distribution offers the most effective size discrimination between the dimensions of the coiled and extended species. In operation, the feed solution and a sweep solution are pumped tangentially across opposite sides of the membrane in co-current fashion, as shown schematically in Fig. 44 (Zydney and van Reis, 2000; van Reis, 2000). Transmembrane pressure is regulated such that filtration occurs at a rate low enough to prevent the rejected solute from accumulating irreversibly on the membrane surface. To help maintain a constant linear velocity of the feed and sweep streams as they traverse the membrane surface, the flow channels are sized with a progressively diminishing cross section and a progressively enlarged cross section on the feed and permeate sides of the membrane module respectively. The rate of cross-sectional-area change corresponds to the volumetric changes of those streams due to filtration. To conserve buffer consumption, HPTFF systems are preferably

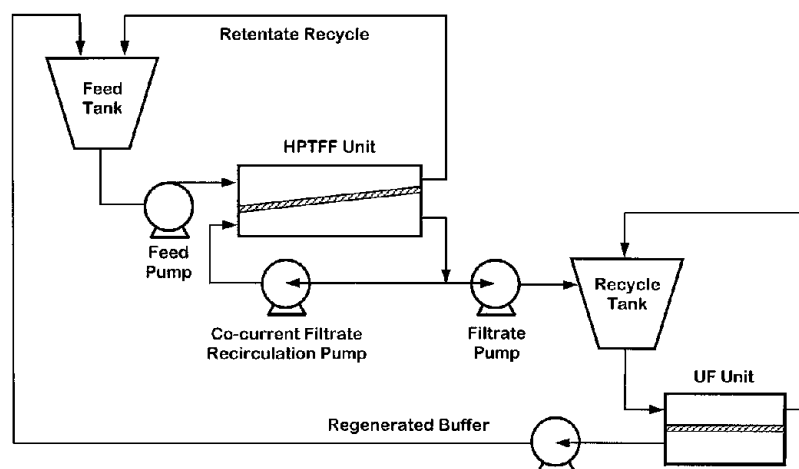


FIGURE 44 Closed-loop cascade HPTFF system operating in diafiltration mode with buffer regeneration by a conventional ultrafiltration unit. [Adapted from Zydney, A. L. and van Reis, R. (2001). In "Membrane Separation in Biotechnology" (W. H. Wong, ed.), Marcel Dekker, New York.]

operated in diafiltration mode, using a second UF stage to regenerate buffer solution from the permeate, as shown in the figure.

Studies show that HPTFF can be used to purify bovine serum albumin (BSA; MW = 68,000 daltons) by removing hemoglobin (Hb; MW = 67,000 daltons) as an impurity. Despite their almost identical molecular weight, hemoglobin exhibits a strong negative charge at pH 7 while BSA is more neutral. By using a negatively charged membrane and adjusting the feed solution to neutral pH, electrostatic repulsion rejects the hemoglobin almost completely. A purification factor of 100 for BSA was achieved. Even more remarkably, BSA may be separated from an antigen-binding fragment (Fab; MW = 45,000) by a purification factor of over 800.

C. Membrane-Based Affinity Separation

Biospecific recognition is among nature's most selective mechanisms. It is the basis of immune responses and the myriad interactions in living organisms. In certain proteins, the combination of amino acid sequence and spatial configuration permits stable binding only with a unique species of complementary functionality and shape. This ligand–ligate recognition and attachment, such as that between an antigen and a monoclonal antibody, or that between specific chemical dyes and proteins, is the principle underlying affinity separations ("Improved tools," 2000).

Commercial exploitation of affinity separation occurred first in the form of column chromatography. Ligands are attached to various passive matrices such as crosslinked cellulosic gel particles. When a solution containing a target protein flows through a packed bed of these gel particles, the target protein attaches to the ligand. Following

a wash step to remove indigenous residues from the column, the protein is detached from the ligand by means of an appropriate buffer solution and recovered free of contaminants. While this method is highly specific, the large pressure drop typical of packed beds also limits the throughput rates achievable. It is also difficult to project large column performance based on the behavior of small systems. Another problem with scale-up is the extraordinary costs associated with populating a large-scale column with affinity ligands, and the correspondingly high risk of loss associated with process upsets. A production-scale batchwise protein separation by column chromatography can require hours or days.

A membrane analog of the affinity column is shown in Fig. 45. Ligand is attached to the internal surfaces of a microporous membrane, which can be thought of as a very wide but very thin (on the order of 10–100 μm) column. Pressure drop is kept low by using a microporous membrane of sufficiently large pore size not to cause separation by physical retention. Target protein is captured when the feed solution flows through the membrane, quickly occupying the limited ligand capacity offered by the membrane matrix. Following a rinse cycle to remove extraneous species held in the membrane, the target protein is released by dissociative elution. Finally, the membrane is regenerated to prepare it for the next cycle of capture/release. Since the hold-up volume is very small, all flow cycles can be quite short; the entire purification sequence can be completed on the order of minutes. By repeating the cycle many times automatically, a small quantity of ligand has the cumulative capacity to harvest the product from a large volume of feed material.

A key feature of affinity membrane separations is the combination of sieving and selective adsorption

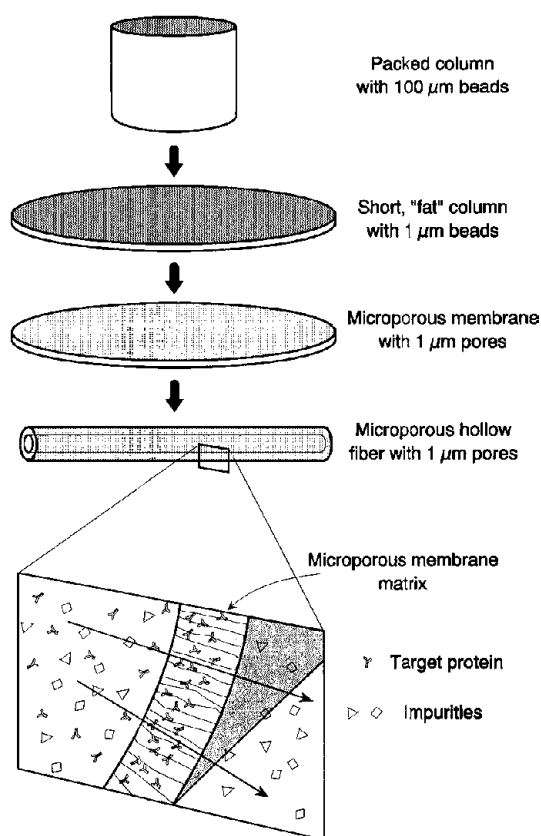


FIGURE 45 Affinity membrane separation of proteins represents the most advanced support for affinity sites.

mechanisms. Both mechanisms could complement each other in enhancing the purity of the product if the separation process is optimized. Typically, the performance of an affinity membrane system is governed by several independent factors: fluid dynamics, ligand–ligate interaction, membrane morphology and ligand distribution, the configuration of the device, and operating conditions. It is possible to achieve very high resolving power using affinity membranes in a stack configuration provided flow through such a stack is controlled to simulate plug flow.

Despite these advantages, membrane affinity processes have not been widely adopted commercially to date for several reasons: affinity chromatography offers a familiar, high-resolution approach to separating complex protein mixtures to the biotechnology industry, in which changes even to a comparable membrane process would require complex and expensive revalidation; scale-up from the laboratory to the process level also requires a series of well-engineered devices and system support, both of which are still at the development stage for affinity membranes.

A comprehensive review of affinity membrane separations (Klein, 2000) traces the development of this technol-

ogy from the perspective of membrane materials, chemical and biochemical functionalization strategies, module design, and engineering considerations. Some applications in biotechnology are mentioned peripherally.

Ion exchange membranes exhibit certain attributes of affinity membranes with respect to bioseparations. Since proteins and other biomolecules carry characteristic charges, it is possible to manipulate the acid–base balance of the environment to cause the species in solution to bind to the membrane, or to release them by shifting away from those conditions. Indeed, ion exchange membranes may be deployed in ways analogous to those for ion exchange chromatography. Engineering considerations similar to those for affinity membranes also apply to ion exchange membrane systems. It is useful to consider all membrane processes, whether they operate largely by sieving or by reversible adsorption, as complementary unit operations in the overall manufacturing scheme and optimized accordingly.

D. Membrane Bioreactors

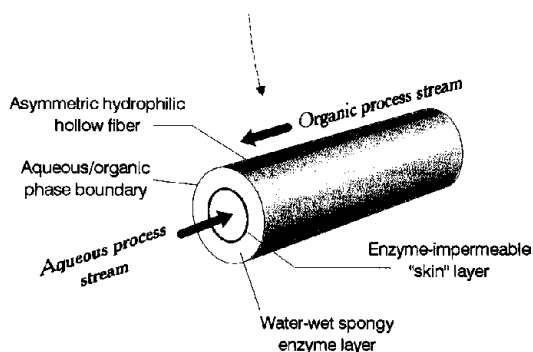
Harnessing biochemical conversions to yield valuable products requires careful control of the environment such that the viability of the catalytic microorganism, enzyme, or cell culture is sustained, plus balanced supplies of reactants and products. Conventional stirred-tank reactors represent the standard configuration of this unit operation, but there are some advantages to using membranes for the same purpose while overcoming some shortcomings of conventional reactor design. As a physical substrate, membranes confine the catalytic species at high packing density and help organize the supply and retrieval of reactants and products, respectively, into separate streams. In addition, the permselective properties of the membrane permit both the conversion and separation functions to be integrated into a single device.

Fermentation is typically conducted in dilute suspension culture. The low concentration in such systems limits reaction efficiency, and the presence of particulate and colloidal solids poses problems for product recovery and purification. By circulating the fermentation broth through an ultrafiltration system, it is possible to recover product continuously as they are generated while minimizing loss of enzyme or cells and keeping product concentration in the bioreactor below the self-inhibition level for the biocatalyst. This process is referred to as perfusion. As the ultrafiltration unit is part of the production process, the entire system is often considered a membrane reactor.

A more appropriate definition of membrane bioreactors confines them to devices in which biochemical conversion actually occurs. For example, an enzyme may be immobilized in the membrane by physical sorption or covalent

attachment to the membrane matrix, or simply confined in solution form in the pores of a membrane and bound externally by immiscible phases. Substrate reaching the enzyme by convection or diffusion is converted to product. Alternatively, the enzyme and the substrate solution may be placed on opposite sides of the membrane so that the reaction may occur at the membrane interface. In either case, the fermentation vessel is eliminated, as is the need to process large volumes of dilute broth at the end of the reaction. Hollow fibers are the preferred geometry for membrane reactors because the biocatalysts can be confined on the shell side, lumen side, or within the macroporous fiber walls, and because circulation on both sides of hollow fibers is convenient.

Attempts at using membrane bioreactors at the process scale have been successful to different extents. One early example of a commercial-scale operation was the enzymatic resolution of optical isomers to produce pharmaceutical intermediates developed in the mid-1980s. By incorporating a stereoselective enzyme in the microporous structure of a membrane, then supplying the racemic substrate on one side of the membrane, one of the isomers is converted by the enzyme and diffuses to the opposite side of the membrane where it can be recovered. The membrane serves several functions simultaneously, as shown in Fig. 46: containing the enzyme, offering a stable interface for contacting the aqueous and organic phases, and removing inhibitory products continuously during the reaction. This elegant process was briefly commercialized to resolve an intermediate to the antihypertensive drug diltiazem (Lopez and Matson, 1997). However, the bioreactor approach was rendered obsolete with the advent of



TYPICAL REACTION:

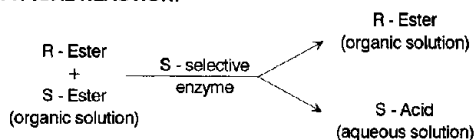


FIGURE 46 Membrane bioreactor in a multiphasic configuration with reversible enzyme containment.

asymmetric synthesis, which directly yields the desired single isomer. Some work on racemic resolution continues today with an emphasis on the production of specialty chemicals.

Other applications of membrane bioreactors proved more sustainable. The first utilizes membranes as an adjunct to conventional bioreactors rather than their replacement. For example, as bioprocesses are scaled up to meet increasing demands of genetically engineered products, every component in a bioreactor needs to be operated as efficiently as possible. For mammalian cell cultures, high cell densities can be reached by replacing traditional methods of gas exchange (e.g., direct sparging, surface or head-space aeration) with gas exchange membranes made from silicone rubber or microporous polytetrafluoroethylene, both of which operate under bubble-free conditions that are much less disruptive to the cell culture. In a similar approach, a “dialysis fermentation” system is assembled by immersing tubular dialysis membrane in a conventional stirred-tank fermenter (“Improved tools,” 2000). Nutrient solution containing glucose and amino acids, for example, is circulated through the dialysis tubing to supply the cells in the fermenter; the same stream carries away spent media containing lactates and ammonia. In this way, cell densities several times higher than those in a conventional fermenter may be reached, resulting in correspondingly higher product yield. Another benefit is that essentially no fermentation product is lost to the nutrient/waste stream because dialysis membranes are only permeable to low-molecular-weight compounds but not to the biomolecules of interest. In a third example, monoclonal antibodies are produced by culturing hybridoma cells in a bioreactor chamber equipped with two membrane systems: a first hollow fiber system circulating nutrient solutions and a second, flat-sheet membrane forming the gas exchange interface to supply oxygen to the cells, as shown in Fig. 47.

Another noteworthy application area for membrane bioreactors is in wastewater or industrial effluent treatment. Since the early to middle 1990s, there has been a resurgence of interest in deploying membrane systems for waste remediation, including oily metal finishing wastes. A combination of increasingly stringent environmental protection regulations, the energy-efficient character of membrane processes, and the ease with which membrane systems may be adapted to different scale operations, have made it attractive to design and operate customized microbial/enzymatic digestion systems using membranes to contain and compartmentalize the biochemical reaction. (Stephenson, Brindle, Judd, and Jefferson, 2000). Much current effort is focused on managing fouling and sustaining reactor productivity. These applications further illustrate a gradual transition in this field. Where membrane bioreactors were once used almost

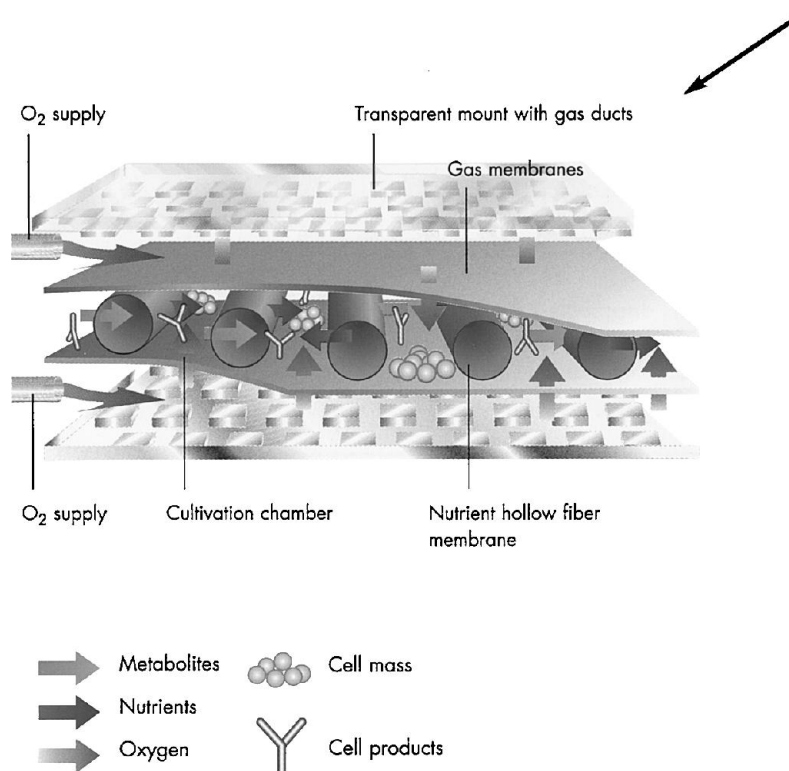


FIGURE 47 Schematic representation of the complex multifunctions enabled in a dual membrane bioreactor for hybridoma cell culturing (Integra Biosciences AG, Wallisellen, Switzerland).

exclusively for producing high-value compounds—such as pharmaceuticals—whose market potential can support the cost of developing new processes, their use has expanded to include processing of lower-value commodities and even wastes through creative integration of proven, increasingly reliable membrane unit operations.

VII. BIOMEDICAL APPLICATIONS

The use of synthetic membranes for medical therapy began with artificial kidney systems. For the better part of the 20th century, hemodialysis has been used as an important life-support process for patients with renal dysfunction. Today this procedure is practiced worldwide, representing the single largest commercial membrane application on the basis of consumption and revenue. Other blood treatment processes involving membranes have also been developed over time, such as hemofiltration, plasmapheresis, blood oxygenation, and various extracorporeal therapies. In parallel, artificial organ concepts have been reduced to practice that include synthetic membranes as key components; some are nearing completion of clinical trial and poised to begin benefiting patients.

A. Hemodialysis, Hemofiltration

A major physiological function of the kidney is to remove toxic metabolic wastes and excess fluids from the blood stream. This function may be impaired through chronic degeneration or as a result of injury. Hemodialysis is the most commonly prescribed means of blood purification for end-stage renal disease. The worldwide dialysis patient population approached 1 million at the end of the 20th century.

In hemodialysis, blood from the patient flows on one side of a membrane and a specially prepared dialysis solution is fed to the other side. Waste material in the blood such as urea, excess acids, and electrolytes diffuse into the dialysate; the blood is then returned to the patient, as shown in Fig. 48. A patient typically undergoes dialysis three times per week in sessions lasting several hours each. Modern dialysis systems combine sophisticated monitoring and control functions to ensure safe operation. Regenerated cellulose was the first material used in hemodialysis membranes because of its biocompatibility and low cost; it remains the most popular choice. Subsequently, high-permeability dialysis membranes derived from cellulose esters, modified polysulfone, or polyacrylonitrile copolymers have also gained wide acceptance because of the shorter sessions they make possible.

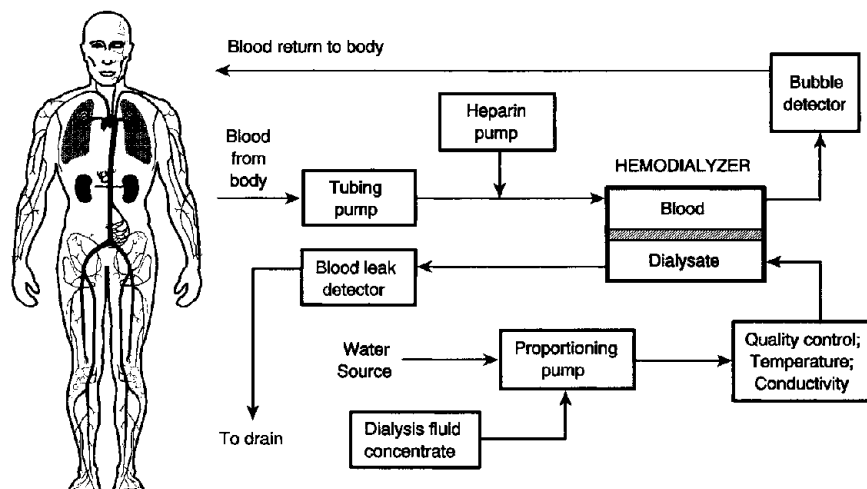


FIGURE 48 Schematic diagram of a hemodialysis system.

Hemofiltration (also referred to as hemodiafiltration) operates by ultrafiltering the blood to remove water and small molecules from plasma. As the process operates by forced convective flow across the membrane instead of diffusion of the solutes alone, a dialysate solution is not provided. Instead, the volume lost by filtration is replaced with a substitution fluid. Because of its ability to remove waste materials rapidly, hemofiltration has been prescribed to treat acute kidney failures associated with intoxication by poison or drugs. It also holds promise as the basic process by which wearable or implantable replacements for the kidney may be operated. As routine therapy, however, hemofiltration may have lost its rate advantage over dialysis after high-flux membranes for that procedure appeared. Also, the need to supply sterile solutions for plasma replenishment in hemofiltration adds significantly to the treatment cost. On the other hand, hemofiltration has found increasing use as an adjunct to surgery, where blood volume is increased prior to surgery by infusion of fluids, then later restored by ultrafiltering excess fluid after the procedure is completed.

B. Plasmapheresis

Plasmapheresis, also known as plasma exchange, is a process developed in the 1970s and now used routinely to treat autoimmune disorders. In such diseases the patient's immune system mistakenly targets tissues in the body for attack. Often this occurs through the generation of autoantibodies by abnormal cells, which accumulate in the bloodstream. As an adjunct to using immunosuppressants, which minimizes formation of autoantibodies but have many side effects if employed alone over long periods, plasma containing the autoantibodies may be removed and replaced by a substitute fluid. Plasmapheresis accom-

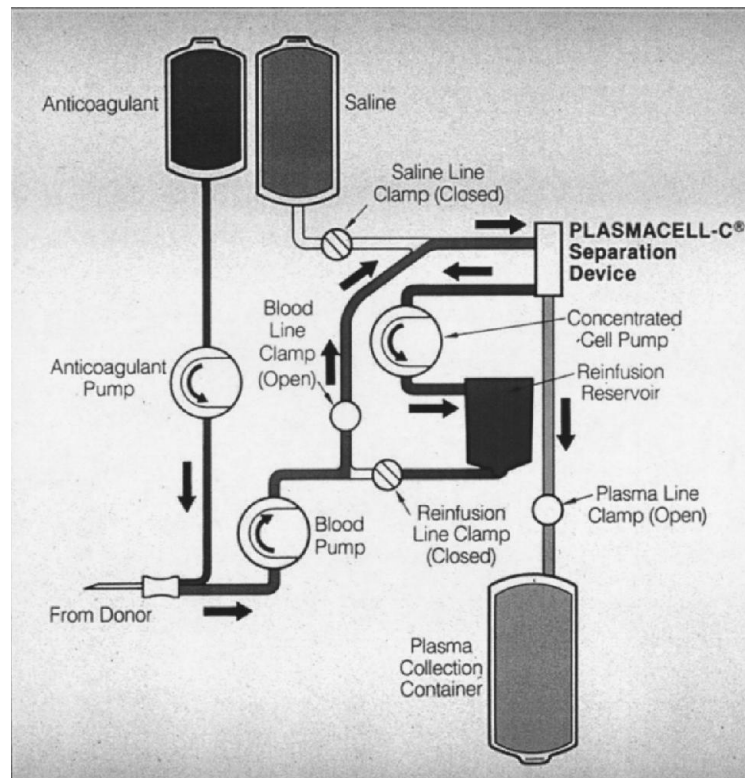
plishes this by separating the blood cells from the plasma with a membrane; the plasma is discarded and the blood cells are returned to the patient. With this approach, significantly lower doses of immunosuppressants are sufficient to manage the condition effectively.

Plasmapheresis typically employs a membrane module of similar configuration as a high-flux hemodialyzer. Alternatively, a rotating membrane separation element is used in which the tendency of the blood cells to deposit on the membrane surface is counteracted with hydrodynamic lift forces created by the rotation. The membrane element and the associated plasmapheresis circuitry are shown in Fig. 49. Worldwide, about 6 million plasmapheresis procedures are performed annually using this system, making this one of the largest biomedical membrane applications after hemodialysis.

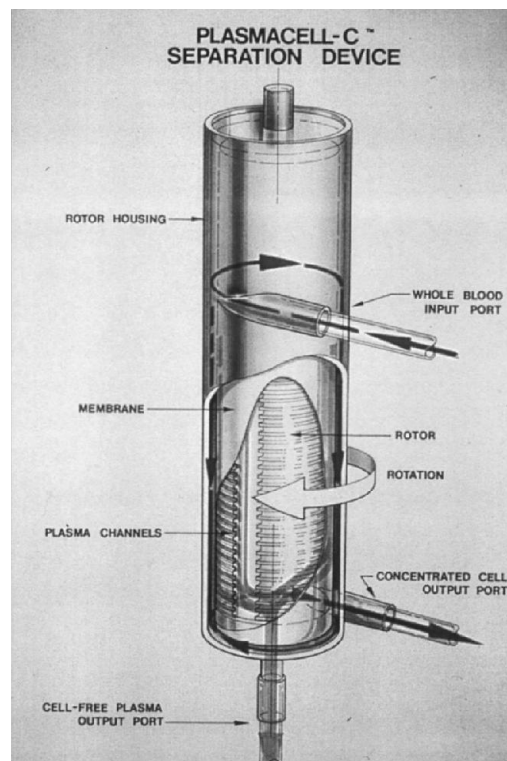
Although proven effective in years of clinical practice, membrane plasmapheresis faces competition from improved centrifuges capable of continuous cell separation. Both approaches are likely to continue playing important roles in the area of extracorporeal therapy.

C. Blood Oxygenation

Blood oxygenation is performed mostly during cardiopulmonary bypass surgery to supplement the reduced capacity of the lung to exchange oxygen and carbon dioxide. Membrane oxygenators containing highly gas-permeable membranes such as silicone rubber are available commercially as an alternative to bubble oxygenators. Oxygen permeating from the gas phase dissolves in the blood stream; device integrity is especially important in this application since leakage of gaseous oxygen into the bloodstream can lead to serious complications.



(a)



(b)

FIGURE 49 (a) Plasmapheresis system and (b) rotating membrane separation device. (Source: Baxter Healthcare Corporation.)

Advances in fabrication technology have made possible the production of very precisely aligned and spaced hollow fibers. Flow across such fiber arrays can thus be regulated to minimize boundary layer effects or blind spots; very high mass transfer efficiencies have been achieved in this manner.

D. Artificial Organs

Membrane-based artificial organs are sophisticated bioreactors. In fact, experience in the area of mammalian cell culture probably first inspired, and then contributed directly to the design and engineering of devices and systems intended as substitutes for healthy human organs. There is certainly great appeal in being able to encapsulate living organ cells in a synthetic membrane, keeping them viable by providing favorable microenvironments while protecting them from immunological attack by the host, and extracting metabolites with therapeutic functions from the cells.

However, unlike bioreactors described previously, which are used to perform a sequence of well-defined biochemical conversions, artificial organs must provide the highly complex metabolic and endocrine functions of the native organ. This has not yet been accomplished, partly because of the difficulty of duplicating all essential regulatory and feedback mechanisms between an organ and its host, and partly because successful demonstration of artificial organ systems, particularly in humans, is very expensive and subject to close institutional and regulatory scrutiny. Finally, membrane-based artificial organs face competition from other approaches to organ replacement, such as xenotransplantation or regeneration of organs from stem cells.

An early demonstration of the artificial organ concept consisted of sealing a small number of bovine cells in the lumen of a microporous hollow fiber. The cells were selected for their ability to secrete a mixture of analgesic biomolecules. Implanting the hollow fiber in the body of a patient places the encapsulated cells under physiological conditions sufficiently favorable for them to function, yet protected from immunological attack by the host—that is, the patient's own defense mechanism against xenogeneic cells. The cells responded by secreting the expected pain-killing agents to the patient. Considerable progress has been made in the methods of encapsulating cells and the design of membrane materials and structure, and the range of target therapies (Lysaght and Aebischer, 1999; Li, 1998).

This approach continues to stimulate innovations in immunoisolated cell therapies. In a very recent animal study (“Study touts,” 2000), cells capable of producing endostatin were encapsulated individually with an algi-

nate coating, effectively creating a large number of single bioreactors. Upon injection into tumor sites of an animal, the cells released endostatin continually to cause anti-angiogenesis, or starvation of the blood supply to the tumor. The result was dramatic shrinkage of the tumor itself. In this example, the alginate coating on each cell serves all the selective permeation and protective functions of an artificial organ membrane. It is not difficult to envision dramatic progress being made in this area of biomedicine—and the enabling role played by synthetic membrane science and technology.

On the commercial front, an artificial liver system has reached advanced clinical trial stage. Based on pig hepatocytes immobilized in a hollow-fiber membrane module, this system provides temporary life support until a liver from a human donor is available for transplantation (Fig. 50). Also under development is an artificial pancreas intended as a permanent replacement of the native organ (Fig. 51).

E. Controlled Release

Safe and efficient use of many pharmaceuticals and therapeutic agents in general requires that the dosage and delivery rate be precisely regulated. An agent must reach

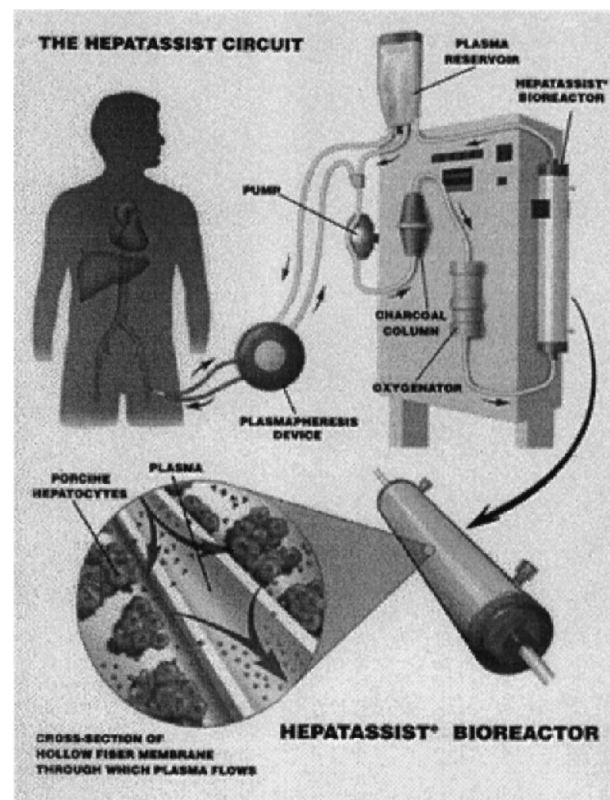


FIGURE 50 Schematic of advanced artificial liver system undergoing clinical trials (Circe Biomedical, Inc., Lexington, MA).

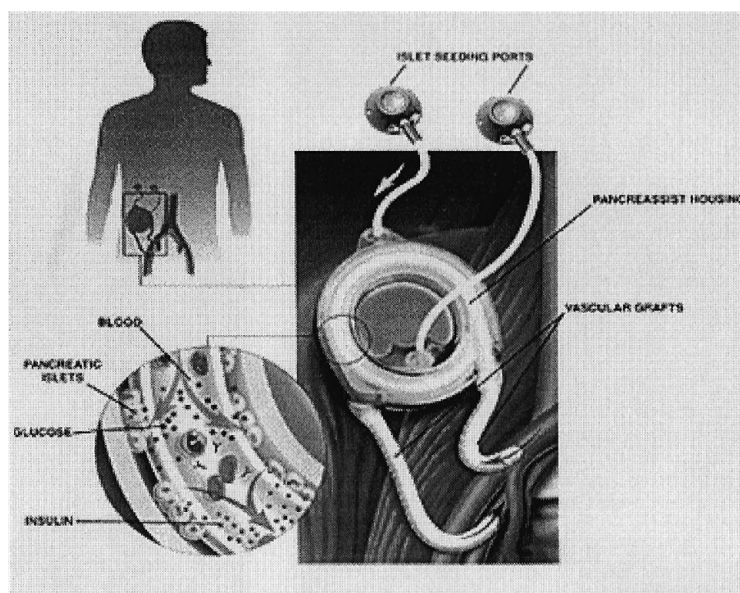


FIGURE 51 Schematic of artificial pancreas under development based on a membrane unit (Circe Biomedical, Inc., Lexington, MA).

certain minimum level to obtain the desired therapeutic effect, but below the point where toxic side effects occur. Once the desired level is reached, it should be maintained over an extended period without further intervention from the user. Controlled release technology based on the use of synthetic membranes is an effective approach to addressing these issues.

1. Passive Controlled Release

In one type of controlled-release device, an active agent is dissolved at its solubility limit in a reservoir surrounded by a membrane. By providing excess agent in the reservoir (e.g., by physical mixing or compounding), a constant permeation driving force is exerted on the membrane. In this way a constant release rate is achieved until the agent concentration drops below saturation. Release rate is determined by the solubility and diffusivity of the agent in the membrane. Tailoring the chemistry and geometry of the device provides different release profiles and capacities. Devices of this type have been developed for administering steroids for fertility control or hormone therapy, or intraocular dispensing of pilocarpine for glaucoma control. For nonmedical applications, similar devices are available in multilayer laminate form for dispensing insect pheromones, insecticides, fragrances, and bactericides.

Sometimes the human skin is used as the rate-controlling membrane for certain pharmaceuticals that are readily absorbed. A removable skin patch containing the drug is applied directly on the body to provide a reservoir of specific capacity; the rate of administration

is controlled by the size of the patch. The patch may be removed to discontinue drug delivery instantly. Devices of this type have been used to administer nitroglycerine for treating angina.

Still another dosage form is microcapsules, which are tiny droplets of an active agent coated with a permselective barrier polymer, all in the form of a suspension. Microcapsules are conveniently formulated for injection. In such cases the barrier polymer is often chosen to be degradable into innocuous products in the body after the active agent has been dispensed.

Osmotic pumps operate by a different principle. In one design, the active agent is mixed with an osmotic agent such as salt or sucrose, and covered with a water-permeable membrane. When the device is immersed in water and the infusion of water generates a pressure inside the membrane, the active agent is forced out of the device through metering holes or capillaries on the surface of the membrane. Alternatively, the active agent is packaged within a sac that separates it from the osmotic agent, which in turn is confined by the water-permeable membrane. The pressure generated by the wetted osmotic agent compresses the sac and exudes the active agent from the device.

2. Active Controlled Release

A novel approach to controlled release integrates biosensing and control functions in a single membrane device. As an example, a membrane responds to changes in glucose level in the body by automatically changing its permeability to insulin. The response mechanism is shown

in Fig. 52. The enzyme glucose oxidase is immobilized within a crosslinked polymer containing amine functional groups. The polymer is normally impermeable to insulin (M.W. = 6000) but permeable to the much smaller glucose molecules (M.W. = 180). Glucose from the body dif-

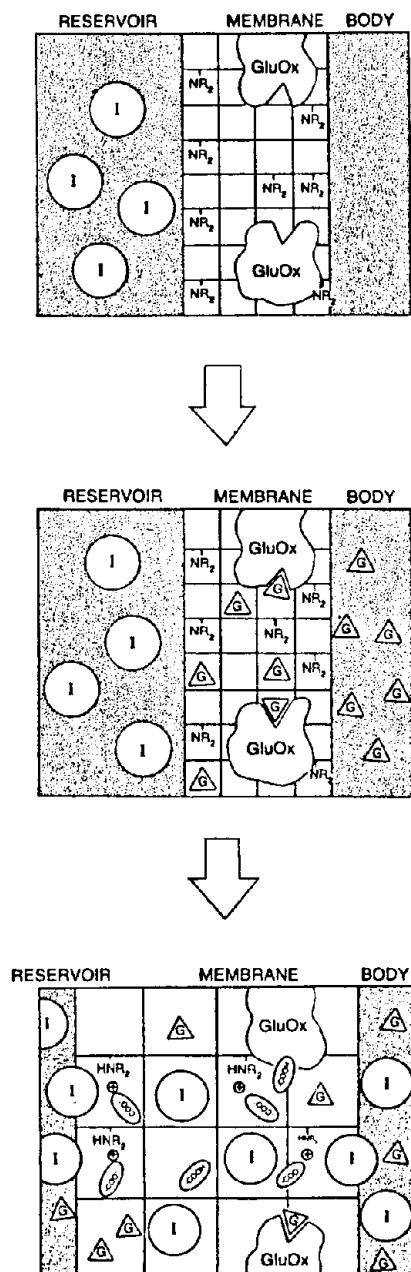


FIGURE 52 A glucose-sensitive membrane with controllable permeability to insulin. (Key: GluOx, glucose oxidase; I, insulin; G, glucose; —COOH, gluconic acid; and —NR₂, tertiary amine functional group in the membrane matrix.) [From Horbett *et al.* (1984). In "Recent Advances in Drug Delivery Systems" (J. M. Anderson and S. W. Kim, eds.), Plenum, New York.]

fuses into the membrane, where it is converted gluconic acid. Dissociation of the acid protonates the tertiary amine groups on the polymer network, and causes it to swell and become more permeable to insulin, which is released into the bloodstream. When the glucose level drops, the enzymatic reaction subsides, and the membrane network contracts to stop passage of insulin. In this case the synthetic membrane closely mimics the complex functional response of a biological membrane. The implications of this development go beyond the obvious application to the treatment of diabetes. The same principle of using biochemical signals to modulate membrane permeability can be applied to the design of other self-regulating sensors and control systems.

VIII. MEMBRANE SENSORS

Two attributes of synthetic membranes are often applied to the design of analytical devices: as a selective barrier, and as a substrate in which chemical or biochemical reactions are performed. In many cases, the membrane helps translate the activity of specific analytes into easily measurable quantities such as electrical potentials or spectrophotometric absorption.

Membrane-based analytical devices that generate electrical responses (current or potential) are referred to as membrane electrodes. The most common membrane electrodes are the ion-selective electrodes (ISE) used for measuring the activity of ions in solution. A typical ISE consists of a reference electrode surrounded by a standard electrolyte solution; an ion-specific membrane separates this "internal solution" from the solution to be analyzed. Only ions to which the membrane is permeable can reach the internal solution to alter the electrochemical balance to generate a signal. The pH electrode is an ISE in which the membrane is a proton-permeable glass. To measure other ions, various membrane materials are designed to provide permeation pathways for the analyte ions. These include single crystals of insoluble inorganic salts, ion-exchange resins dispersed in inert matrices, polymeric ion-exchange membranes, homogeneous polymer films (sometimes swollen with a solvent), and liquid membranes containing carriers or complexing agents. All of these devices make use of the membrane as a selective barrier, and are referred to as primary electrodes.

A sensitized electrode is a composite device that combines a membrane reactor and a primary electrode. The membrane reactor converts a specific analyte into products that are measurable by the primary electrode. Membrane reactors containing immobilized enzymes, cells, or neutral carriers are capable of very selective conversion of sugars, amino acids, organic acids, and alcohols, and

even inorganic ions into several metabolites—oxygen, carbon dioxide, hydrogen ions, and ammonia—that are detectable with primary membrane electrodes. Integrating reactive and sensing membrane components significantly broadens the range of substances that can be analyzed electrochemically.

A family of specialized membrane devices allows complex clinical assays to be performed photochemically. They all contain a membrane matrix with specific reactive moieties. When an analyte containing a target species is introduced, reaction products are formed that may be detected spectrophotometrically. The membrane supplies a large functional surface area per unit device volume and hence high sensitivity. Figure 53 depicts the general structure of devices of this type designed for clinical analysis. These are thin multilayer composites activated by applying a small quantity of liquid clinical sample. A microporous membrane at the top surface distributes the fluid uniformly to the layers beneath, where various physical and chemical reactions occur to form chromophoric products. The quantity of those products is measured spectrophotometrically from the lower surface of the slide.

Different laminated structures are constructed by arranging the functional layers according to the requirements of the chemical or biochemical assay. Commercially available assays include those for creatinine, albumin, amylase, bilirubin, cholesterol, triglycerides, and certain alkali and alkaline earth metals.

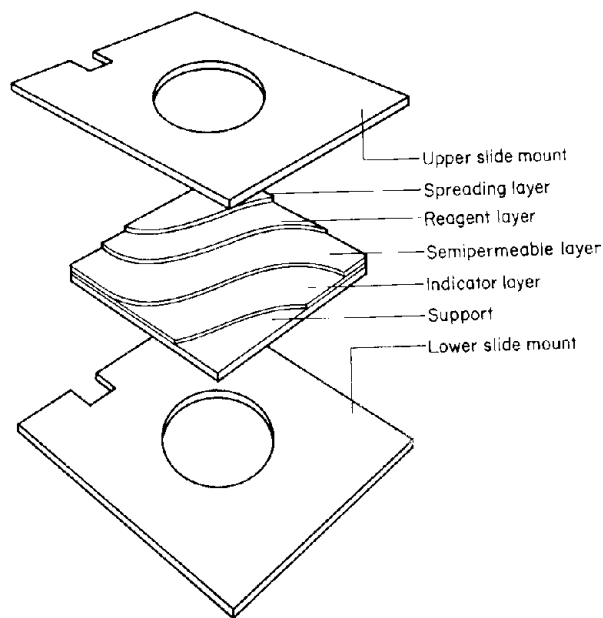


FIGURE 53 Schematic diagram of a photochemical membrane composite for clinical diagnosis. (Eastman Kodak Company.)

SEE ALSO THE FOLLOWING ARTICLES

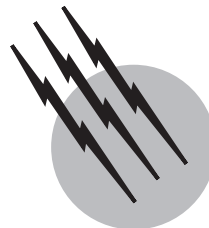
BIOMATERIALS, SYNTHESIS, FABRICATION, AND APPLICATIONS • BIOREACTORS • DISTILLATION • ELECTROCHEMICAL ENGINEERING • FLUID DYNAMICS • MEMBRANE STRUCTURE • MEMBRANES, SYNTHETIC (CHEMISTRY) • MOLECULAR HYDRODYNAMICS • NANOSTRUCTURED MATERIALS, CHEMISTRY OF • PHARMACEUTICALS, CONTROLLED RELEASE OF • SOLVENT EXTRACTION • WASTEWATER TREATMENT AND WATER RECLAMATION

BIBLIOGRAPHY

- Alentiev, A. Yu., Loza, K. A., and Yampol'skii, Y. P. (2000). "Development of methods for prediction of gas permeation parameters of glassy polymers: Polyimides as alternating copolymers," *J. Membrane Sci.* **167**, 91.
- Anand, M., Langsam, M., Rao, M. B., and Sircar, S. (1997). "Multicomponent gas separation by selective surface flow (SSF) and poly-trimethylsilylpropyne (PTMSP) membranes," *J. Membrane Sci.* **123**, 17.
- Araki, T., and Tsukube, H., eds. (1990). "Liquid Membranes: Chemical Applications," CRC Press, Boca Raton, FL.
- Aranha, H. (2001, January). "Viral clearance strategies for biopharmaceutical safety, Part 1: general considerations," *BioPharm*, pp. 28–35.
- Aranha, H. (2001, February). "Viral clearance strategies for biopharmaceutical safety, Part 2: Filtration for viral clearance," *BioPharm*, pp. 32–43.
- Baker, R. W. (1999, December). "Membrane Technology and Applications," McGraw-Hill, New York.
- Baker, R. W., Cussler, E. L., Eykamp, W., Koros, W. J., Riley, R. L., and Strathmann, H. (1991). "Membrane Separation Systems, Recent Developments and Future Directions," Noyes Data Corporation, Park Ridge, NJ.
- Baker, R. W., et al. (1998). "The design of membrane–gas separation systems," *J. Membrane Sci.* **151**(1), 55–62.
- Baker, R. W., et al. (2000, December). "The design of membrane–gas separation systems," *Chem. Eng. Prog.* pp. 51–57.
- Balaban, M., ed. (1991, May). "Desalination and Water Re-Use: Proceedings of the 12th International Symposium: Economics Membrane Distillation Membrane Processes Evaporative Processes," Hemisphere, Bristol, PA.
- Balachandran, U., Dusek, J. T., Maiya, P. S., Ma, B., Mieville, R. L., Kleefisch, M. S., and Udovich, C. A. (1997). "Ceramic membrane reactor for converting methane to syngas," *Catalysis Today* **36**(3), 265.
- Balachandran, U., Kleefisch, M. S., Kobylinski, T. P., Morissette, S. L., and Pei, S. (1996). "Oxygen ion-conducting dense ceramic," U.S. Patent 5580497.
- Bartsch, R. A., and Way, J. D., eds. (1996). "Chemical Separations With Liquid Membranes, ACS Symposium Series, No 642," American Chemical Society, Washington, DC.
- Belfort, G., Davis, R. H., and Zydney, A. L. (1994). "The behavior of suspensions and macromolecular solutions in crossflow microfiltration," *J. Membr. Sci.* **96**, 1.
- Bessarabov, D. (1999). "Membrane gas-separation technology in the petrochemical industry," *Membrane Technol.* **107**, 9.
- Burggraaf, A. J., and Cot, H., eds. (1996). "Fundamentals of Inorganic

- Membrane Science and Technology (Membrane Science and Technology, Vol. 4), Elsevier Science, Amsterdam.
- Cardew, P. T., and Level, M. S., eds. (1995, January). "Membrane Processes: A Technology Guide," Springer-Verlag, New York.
- Carman, P. C. (1937). "Fluid Flow through Granular Beds," *Trans. Inst. Chem. Eng. London* **15**, 150.
- Cheryan, M. (1986). "Ultrafiltration Handbook," Technomic Publishing Co., Lancaster, PA, pp. 231–233.
- Coulson, J. M. (1949). "The flow of fluids through granular beds: Effect of particle shape and voids in streamline flow," *Trans. Inst. Chem. Eng.* **27**, 237.
- Crank, J. (1975). "The Mathematics of Diffusion," 2nd ed., Oxford University Press, Oxford.
- Cussler, E. L. (1994). In "Polymeric Gas Separation Membranes" (D. R. Paul and Y. P. Yampol'skii, eds.), pp. 273–300, CRC Press, Boca Raton, FL.
- Dalvie, S. K., and Baltus, R. E. (1992). "Transport studies with porous alumina membranes," *J. Membr. Sci.* **71**, 247.
- Datta, R., and Tsai, S.-P. (1998, March 3). "Esterification of fermentation-derived acids via pervaporation," U.S. Patent 5723639.
- Deen, W. M. (1987). "Hindered transport of large molecules in liquid-filled pores," *AIChE J.* **33**, 1409.
- Fane, A. G., and Radovich, J. M. (1990). In "Separation Processes in Biotechnology" (J. A. Asenjo, ed.), Chap. 8, Marcel Dekker, New York.
- Freeman, B. D. (1999). "Basis of permeability/selectivity tradeoff relations in polymeric gas separation membranes," *Macromolecules* **32**(2), 375–380.
- Gas Processors Association Standard 3132-84.
- Glasstone, S. (1950). "Textbook of Physical Chemistry," D. Van Nostrand, New York.
- Happel, J., and Brenner, H. (1965). "Low Reynolds Number Hydrodynamics with Special Applications to Particulate Media," Prentice-Hall, Englewood Cliffs, NJ.
- Heitner-Wirguin, C. (1996). "Recent advances in perfluorinated ionomer membranes: Structure, properties and applications," *J. Membrane Sci.* **120**, 1.
- Ho, W. S. W. (2000, May 25). "Strontium removal with supported liquid membranes," Paper presented at the 11th Annual NAMS Conference on Membranes, Boulder, Colorado.
- Ho, W. S. S., and Sirkar, K. K. (1992). "Membrane Handbook," Van Nostrand Reinhold, New York.
- Hsieh, H. P., ed. (1996). "Inorganic membranes for separation and reaction (Membrane Science and Technology, Vol. 3)," Elsevier Science, New York.
- "Improved tools for large-scale bioprocessing" (2001, April 11). *Genet. Eng. News* **20**, 11.
- Jakobs, E. (1996). "Modification and characterization of mass-transfer properties of gamma-alumina membranes," Ph.D. dissertation, University of Texas at Austin.
- Kesting, R. E., and Fritzsche, A. K. (1993). "Polymeric Gas Separation Membranes," John Wiley & Sons, New York.
- Klein, E. (1991). "Affinity Membranes: Their Chemistry and Performance in Adsorptive Separation Processes," John Wiley & Sons, New York.
- Klein, E. (2000). "Affinity membranes: A ten-year review," *J. Membrane Sci.* **179**, 1.
- Koros, W. J. (1985). "Simplified analysis of gas/polymer selective solubility behavior," *J. Polym. Sci. Polym. Phys. Ed.* **23**, 1611.
- Koros, W. J. (1995). "Membranes: Learning a lesson from nature," *Chem. Eng. Prog.* **91**, 68.
- Koros, W. J., and Hellums, M. W. (1989). In "Concise Encyclopedia of Polymer Science and Engineering," 2nd ed. (J. I. Kroschwitz, ex. ed.), pp. 1211–1219, Wiley-Interscience, New York.
- Koros, W. J., and Mahajan, R. (2000). "Pushing the limits on possibilities for large scale gas separation: Which strategies?" *J. Membr. Sci.* **175**, 181–196.
- Koros, W. J., and Pinnau, I. (1994). In "Polymeric Gas Separation Membranes" (D. R. Paul and Y. P. Yampol'skii, eds.), pp. 209–272, CRC Press, Boca Raton, FL.
- Koros, W. J., Coleman, M. R., and Walker, D. R. B. (1992). "Controlled permeability polymer membranes," *Annu. Rev. Mater. Sci.* **22**, 47.
- Koros, W. J., Fleming, G. K., Jordan, S. M., Kim, T. H., and Hoehn, H. H. (1988). "Polymeric membrane materials for solution-diffusion based permeation separations," *Prog. Polym. Sci.* **13**, 339.
- Lamendolar, M. F., and Tua, A. (1995). "The Malta experience: Desalination of seawater by reverse osmosis," *Desalination and Water Reuse* **50**, 18.
- Lee, E. K. (1993). In "Science for the Food Industry of the 21st Century" (M. Yalpani, ed.), pp. 195–212, ATL Press, Science Publishers, Mt. Prospect, IL.
- Leenaars, A. F. M., and Burggraaf, A. J. (1985). "The preparation and characterization of alumina membranes with ultrafine pores: Part 2. The formation of supported membranes," *J. Coll. Inter. Sci.* **105**, 27.
- Levy, R. V., Phillips, M. W., and Lutz, H. (1998). In "Filtration in the Biopharmaceutical Industry" (T. Meltzer and M. W. Jorntz, eds.), Marcel Dekker, New York.
- Li, R. H. (1998). "Materials for immunisolated cell transplantation," *Adv. Drug Delivery Rev.* **33**, 87–109.
- Lin, Y. S., Wang, W., and Han, J. (1994). "Oxygen permeation through thin mixed-conducting solid oxide membranes," *AIChE J.* **40**(5), 786.
- Lopez, J. L., and Matson, S. L. (1997). "Multiphase/extractive enzyme membrane reactor for production of diltiazem chiral intermediate," *J. Membrane Sci.* **125**(1), 189–211.
- Lysaght, M. J., and Aebischer, P. (1999, April). "Encapsulated cells as therapy," *Sci. Am.* pp. 76–82.
- Ma, Y. H. (1999). "Dense palladium and perovskite membranes and membrane reactors," *MRS Bull.* **24**(3), 46.
- Mahajan, R., and Koros, W. J. (1999, June). "Mixed matrix gas separation membranes," International Congress on Membranes, Toronto Canada.
- Mahajan, R., Zimmerman, C. M., and Koros, W. J. (1997). In "Polymer Membranes for Gas and Vapor Separation, ACS Symposium Series 733" (B. Freeman and I. Pinnau, eds.), pp. 277–286, American Chemical Society, Washington, DC.
- Matsuura, T., and Sourirajan, S. (1981). "Reverse osmosis transport through capillary pores under the influence of surface forces," *Ind. Engr. Chem., Proc. Des. Dev.* **20**, 273.
- McGregor, W. C. (1986). "Membrane Separations in Biotechnology," Marcel Dekker, New York.
- Meindersma, G. W., and Kuczynski, M. (1996). "Implementing membrane technology in the process industry: Problems and opportunities," *J. Membrane Sci.* **113**, 285.
- Meltzer, T., and Jorntz, M. W., eds. (1998). "Filtration in the Biopharmaceutical Industry," Marcel Dekker, New York.
- Merten, U., ed. (1966). "Desalination by Reverse Osmosis," MIT Press, Cambridge, MA.
- Miller, J. R. (1992). "Transport properties of native and chemically modified gamma-alumina membranes," Ph.D. dissertation, University of Texas at Austin.
- Mitchell, B. C., and Deen, W. M. (1986). "Effect of concentration on the rejection coefficients of rigid macromolecules in track-etch membranes," *J. Coll. Inter. Sci.* **113**, 132.
- Morooka, S., and Kusabe, K. (1999). "Microporous inorganic membranes for gas separation," *MRS Bull.* **24**(3), 25.
- Mulder, M. (1996, October). "Basic Principles of Membrane Technology," Kluwer Academic, New York.

- Nakao, S. (1994). "Determination of pore size and pore size distribution 3. Filtration membranes," *J. Membr. Sci.* **96**, 93.
- Nigara, Y., Mizusaki, J., and Ishigame, M. (1995). "Measurement of oxygen permeability in CeO₂ doped CSZ," *Solid State Ionics* **79**, 208.
- Noble, R. D., and Stern, S. A., eds. (1995, January). "Membrane Separations Technology: Principles and Applications, Membrane Science and Technology, Vol. 2," Elsevier Science, New York.
- Park, J. Y., and Paul, D. R. (1997). "Correlation and prediction of gas permeability in glassy polymer membrane materials via a modified free volume based group contribution method," *J. Membrane Sci.* **125**, 23.
- Pinnau, I., ed. (1999). "Polymer Membranes for Gas and Vapor Separation: Chemistry and Materials Science, ACS Symposium Series, 733," American Chemical Society, Washington, DC.
- Pivovar, B. S., Wang, Y., and Cussler, E. L. (1999). "Pervaporation membranes in direct methanol fuel cells," *J. Membrane Sci.* **154**, 155.
- Prasad, R., Notaro, F., and Thompson, D. R. (1994). "Evolution of membranes in commercial air separation," *J. Membrane Sci.* **94**, 225.
- Puri, P. S. (1996). "Gas separation membranes current status," *La Chimica e l'Industria* **78**, 815.
- Rao, M. B., and Sircar, S. (1993). "Nanoporous carbon membrane for gas separation," *Gas Separation and Purification* **7**(4), 279.
- Rao, M. B., and Sircar, S. (1997). "Nanoporous carbon membranes for separation of gas mixtures by selective surface flow," *J. Membrane Sci.* **85**, 253.
- Rezac, M. E., and Schoberl, B. (1999). "Transport and thermal properties of poly (ether imide)/acetylene-terminated monomer blends," *J. Membrane Sci.* **156**, 211.
- Robeson, L. M. (1991). "Correlation of separation factor versus permeability for polymeric membranes," *J. Membrane Sci.* **62**, 165.
- Robeson, L. M., Smith, C. D., and Langsam, M. (1997). "A group contribution approach to predict permeability and permselectivity of aromatic polymers," *J. Membrane Sci.* **132**, 33.
- Sakai, K. (1994). "Determination of pore size and pore size distribution. 2. Dialysis membranes," *J. Membr. Sci.* **96**, 93.
- Scott, J., ed. (1981). "Desalination of Seawater by Reverse Osmosis," Noyes Data Corporation, Park Ridge, NJ.
- Segre, G., and Silberberg, A. (1962). *J. Fluid Mech.* **14**, 13.
- Singh, A., and Koros, W. J. (1996). "Significance of entropic selectivity for advanced gas separation membranes," *Ind. Eng. Chem. Res.* **35**, 1231.
- Singh, R. (1999). "Will developing countries spur fuel cell surge?" *Chem. Eng. Prog.* **95**(3), 59.
- Spillman, R. W. (1989). "Economics of gas separation membranes," *Chem. Eng. Prog.* **85**, 41.
- Staudt-Bickel, C., and Koros, W. J. (1999). "Improvement of CO₂/CH₄ separation characteristics of polyimides by chemical crosslinking," *J. Membrane Sci.* **155**, 145.
- Stephenson, T., Brindle, K., Judd, S., and Jefferson, B. (2000, July). "Membrane Bioreactors for Wastewater Treatment," IWA Publishing.
- Stern, S. A., and Koros, W. J. (2000). "Separation of gas mixtures with polymer membranes: A brief overview," *Chimie Nouvelle* **18**(72), 3201–3215.
- Story, B. J., and Koros, W. J. (1991). "Sorption of carbon dioxide/methane mixtures in poly(phenylene oxide) and a carboxylated derivative," *J. Appl. Polym. Sci.* **42**, 2613.
- "Study touts injection of cancer drug" (2000, December 30). *Boston Globe*.
- Thornborough, J. R., ed. (1994, November). "Membrane Function: Membrane Structure and Function, Membrane Transport of Non-electrolytes, Membrane Transport of Electrolytes," McGraw-Hill, New York.
- Tsapatis, M., and Gavalas, G. R. (1999). "Synthesis of porous inorganic membranes," *MRS Bull.* **24**(3), 30.
- Turner, M. K. (1991). "Effective Industrial Membrane Processes: Benefits and Opportunities," Elsevier Applied Science, London.
- van Reis, R. (2000, April 25). "Tangential-flow filtration system," U.S. Patent 6,054,051.
- Way, J. D., ed. (1996, May). "Chemical Separations with Liquid Membranes," American Chemical Society, Washington, DC.
- Way, J. D., and Noble, R. D. (1992). In "Membrane Handbook" (W. S. W. Ho and K. K. Sircar, eds.), pp. 833–866, Chapman and Hall, New York.
- Wood, B. (1968). "Dehydrogenation of cyclohexane on a hydrogen-porous membrane," *J. Catal.* **11**(1), 30.
- Yeagle, P., ed. (1992). "The Structure of Biological Membranes," CRC Press, Boca Raton, FL.
- Zeman, L. J., and Zydney, A. L. (1996). "Microfiltration and Ultrafiltration: Principles and Applications," Marcel Dekker, New York.
- Zolanz, R. R., and Fleming, G. K. (1992). In "Membrane Handbook" (W. S. W. Ho and K. K. Sircar, eds.), pp. 54–77, Chapman and Hall, New York.
- Zydney, A. L., and van Reis, R. (2001). In "Membrane Separations in Biotechnology" (W. H. Wong, ed.), Marcel Dekker, New York.



Metalorganic Chemical Vapor Deposition (MOCVD)

Russell D. Dupuis

University of Texas at Austin

- I. Summary of the Metalorganic Chemical Vapor Deposition Process
- II. Properties of Common Metalorganics and Hydrides Used for MOCVD
- III. Growth of III–V Compound Semiconductors by MOCVD
- IV. Some Representative Other Materials Grown by MOCVD
- V. Other Developments in MOCVD
- VI. Future Vision
- VII. Summary and Conclusions

GLOSSARY

Chemical vapor deposition (CVD) The deposition of materials using one or more chemically driven processes that employ vapor-phase transport of the reagents to the deposition zone.

Epitaxy The deposition of a single-crystal film of a material upon a template of atoms provided by the surface of a crystalline solid called the “*substrate*.” Such a film is termed an “*epitaxial layer*.” If the film and substrate are composed of materials having the same lattice parameter, the film is “*homoepitaxial*,” and if the film and substrate are formed from materials with different lattice parameters, the film is “*heteroepitaxial*.”

Five-to-three ratio ([V]/[III] or simply V/III) The ratio of the sum of the partial pressures of all precursors for Column V species to the sum of the partial pressures of all of the Column III precursors in a growth process. Normally, the V/III ratio for MOCVD is much greater than 1.

Heterogeneous chemical reactions Chemical reactions that occur between a precursor in one phase at the interface between this phase and another phase (e.g., a vapor phase and a solid phase).

Homogeneous chemical reactions Chemical reactions that occur only between precursors in the same phase as the phase of the input precursors. The phase may be a solid, liquid, or gas phase.

Hydride A chemical compound containing only chemical bonds between atoms of an element and hydrogen atoms (e.g., AsH_3).

Metalorganic A compound containing one or more chemical bonds between a metal atom and the carbon atoms of an organic radical, e.g., $(\text{CH}_3)_3\text{Ga}$; these compounds are also known as “*metal alkyls*” or “*organometallics*.”

Precursor A chemical compound that is used as an “input” to a chemical process to produce a desired product. Precursors are often referred to as “*sources*.”

Pressure In most CVD processes, the pressures are described relative to “standard atmospheric pressure” measured at the earth’s surface (~ 14.7 lbs/in² absolute or PSIA), equivalent to 760 Torr, 1000 mbar, or 100 kPa.

Pyrolysis The decomposition of a compound using thermal energy.

Pyrophoric A compound that reacts with air in a way that results in spontaneous ignition.

III–V compound semiconductor A semiconductor that in pure form is composed of a mixture of atoms of one or more elements from Column III and an equal number of atoms of one or more elements from Column V of the Periodic Table. The Column III atoms are arranged on one sublattice and the Column V atoms are on another.

Vapor-phase epitaxy (VPE) An epitaxial growth process that only uses chemical precursors that are delivered to the growing surface in the vapor-phase.

METALORGANIC CHEMICAL VAPOR DEPOSITION (MOCVD) is a process employing the pyrolysis of vapor-phase mixtures of various chemical reagents (i.e., precursors) to form thin solid films of materials as diverse as metals, semiconductors, and insulators. Currently, the primary use of MOCVD is for the growth of crystalline thin films of semiconductors and related materials. The application of MOCVD to the growth of the III–V compound semiconductors will be the primary focus of this article. The metalorganic chemical vapor deposition technology has advanced remarkably since the first report of the growth of semiconductor epitaxial films by H. M. Manasevit in 1968. The first high-performance semiconductor devices realized by MOCVD were AlGaAs/GaAs injection lasers and solar cells reported by R. D. Dupuis *et al.* in 1977. In this past thirty-odd years, MOCVD has been developed for the production of AlGaAs, InAlGaP, InGaAsP, InAlGaP, and a variety other III–V compound semiconductor materials. It is now the dominant technology worldwide for the commercial production of light-emitting diodes, injection lasers, quantum-well lasers, solar cells, photodetectors, heterojunction bipolar transis-

tors, and a variety of other electronic and optoelectronic devices. This paper will review some of the important aspects of this technology.

I. SUMMARY OF THE METALORGANIC CHEMICAL VAPOR DEPOSITION PROCESS

A. Nomenclature

Over the years since 1968, there have been several other names applied to this process, including metal-alkyl vapor phase epitaxy (MAVPE), metalorganic VPE (MOVPE), organometallic CVD (OMCVD), and organometallic VPE (OMVPE). However, Manasevit first used the term “metalorganic” (emphasizing the *metal* component) because that was the common term applied to the “metal alkyl” compounds at this time and “CVD” because he felt that the process could be broadly applied to “chemical vapor deposition” of many different materials, including polycrystalline and amorphous films—the term “vapor-phase epitaxy” is a special case of the more general term “chemical vapor deposition.” Later, the term “organometallic” (emphasizing the *organic* component) came to be applied to these specific metal alkyl compounds by the synthetic chemists studying these materials. This paper will use the more “generic” and historical name “metalorganic chemical vapor deposition” for this process.

B. Historical Development of MOCVD

The MOCVD technology for the growth of III–V compound semiconductors has been extensively developed since its introduction, and has today become the dominant epitaxial materials technology for both research and production of III–V compound semiconductors. Because of the flexibility in the growth process and the materials quality of films produced by MOCVD, many important III–V devices have become commercially viable. The MOCVD epitaxial growth technology as we know it today was first reported in the scientific literature in early 1968 by H. M. Manasevit (North American Rockwell, United States). However, prior to 1967, similar processes and experimental results had been previously described in the patent literature by other groups, e.g., T. R. Scott *et al.* (Standard Telecommunications and Cables, United Kingdom), W. Miederer *et al.* (Siemens, West Germany), and R. A. Ruehrwein (Monsanto, United States). In 1967, Manasevit was primarily interested in technologies for the heteroepitaxial growth of III–Vs on insulating substrates, the analogue of the silicon-on-insulator (SOI) and silicon-on-sapphire (SOS) technology that he had developed earlier.

However, in Manasevit's first paper on MOCVD, concerning the epitaxial growth of GaAs on insulators, the actual epitaxial process was **not even mentioned!**

It is interesting to note that besides MOCVD, many important innovations in III–V compound semiconductor epitaxial growth technologies were first developed in the 1966–1967 time frame. For example, J. J. Tietjen and J. A. Amick (RCA Laboratory, USA) first reported “open-tube” hydride VPE growth of III–Vs in 1966 and H. Rupprecht *et al.* (IBM Laboratory, United States) first reported liquid-phase epitaxial (LPE) growth of ternary alloys of $\text{Al}_x\text{Ga}_{1-x}\text{As}$ in 1967. Also in 1967, J. A. Arthur (Bell Telephone Laboratories, USA) reported the first studies of the properties of Ga and As molecular beams in ultrahigh vacuum—which ultimately lead to growth of GaAs by molecular beam epitaxy (MBE) reported by A. Y. Cho (Bell Telephone Laboratories, USA) in 1970.

Over the next few years after 1968, Manasevit and coworkers explored the growth of various III–V, II–VI, and IV–VI compounds by MOCVD. Manasevit concentrated on the growth of thin semiconductor films on various *insulating oxide substrates* including sapphire, spinel, and beryllium oxides. Much of Manasevit's work was “proof-of-concept” growth studies on insulators. The impurity content of these films was relatively high for several reasons. First, the purity of the metal alkyl sources was still very far behind that of other precursors used for III–V epitaxial growth, especially compared to the pure metals and the metal halides. Contributions to the impurity concentrations were also made by the hydrides. Furthermore, the MOCVD process is very sensitive to oxygen (more so than LPE and VPE) and the quality of the films is degraded when small oxygen leaks exist in the reactor system. Oxygen incorporation also contributes to excessive C incorporation. Given the state-of-the-art in reactor system design and construction in the late 1960s and early 1970s, this oxygen sensitivity created serious problems, especially for the growth of Al-containing alloys. These combined effects led to low carrier mobilities, high background impurity concentrations, poor surface morphologies, and generally low photoluminescence efficiencies compared to those achieved by other competing and more well-developed III–V materials technologies, e.g., LPE and VPE. While Manasevit and other workers studied the growth of III–Vs by this process in the early and middle 1970s, they were unable to demonstrate materials quality comparable to that of other III–V epitaxial technologies such as liquid-phase epitaxy and halogen- and hydride-based vapor-phase epitaxy.

Since the early 1960s and into the middle and late 1970s, various other III–V materials technologies had been developed, and had come to dominate the research and production efforts worldwide, including (1) VPE using Column V

halides (e.g., AsCl_3) and Column III metals; (2) VPE using Column V hydride sources (e.g., AsH_3) and Column III trichlorides, e.g., GaCl_3 ; (3) LPE using Column III metal solutions (e.g., Ga melts with GaAs source material); (4) MBE using pure elemental sources (e.g., Ga and As). Already in 1968, when Manasevit's paper first appeared, the VPE and LPE technologies were proven for the growth of a variety of III–V “high-performance” devices. By 1973, hydride VPE dominated the production of GaAsP light-emitting diodes (LEDs) and halide VPE dominated the production of high-purity GaAs for electronic devices. LPE was the dominant technology for many III–V compounds, especially Al-containing devices, including Al-GaAs LEDs, lasers, solar cells, and other heterojunction devices. By 1975, MBE was being actively researched by a few groups, particularly at Bell Laboratories and IBM Research Laboratory. Consequently, there was not much interest in MOCVD—it was viewed as just “another” III–V materials technology—and the materials results seemed to be much worse than those achieved by the other III–V epitaxial growth technologies.

In 1977 R. D. Dupuis *et al.* reported high-performance AlGaAs/GaAs solar cells and injection lasers grown by MOCVD, showing that this technology could perform at levels equal the other III–V materials technologies. In 1978, they reported the first quantum-well semiconductor injection lasers operating continuously at 300 K, clearly showing that the performance of MOCVD-grown devices could, in fact, exceed that of alternate materials technologies. These results caused many groups to reconsider the exploration of MOCVD materials technology, resulting in a rapid increase in the rate of publication of research papers on this topic and its development as a production process for III–V epitaxial films.

C. General Description of the MOCVD Process

The MOCVD process (as applied to the growth of III–V compound semiconductors) generally employs vapor-phase mixtures of Column III metalorganic and Column V hydride sources (precursors) in a carrier gas and is carried out in an open-tube process chamber. In some cases, one or more of the Column V precursors may also be a metalorganic source. The carrier gases are typically purified H_2 or N_2 . The input gas mixtures are heated above $\sim 350^\circ\text{C}$ using a heating system, which provides thermal energy to the growth surface of the substrate. The thermal energy source is most often radio-frequency (RF) induction, electrical resistance, or optical infrared (IR) heating systems. The process chamber total pressure during growth is typically in the range 20–760 Torr (26–1000 mbar, 2.6–100 kPa). The process or “reactor” chamber is usually composed of quartz or stainless steel.

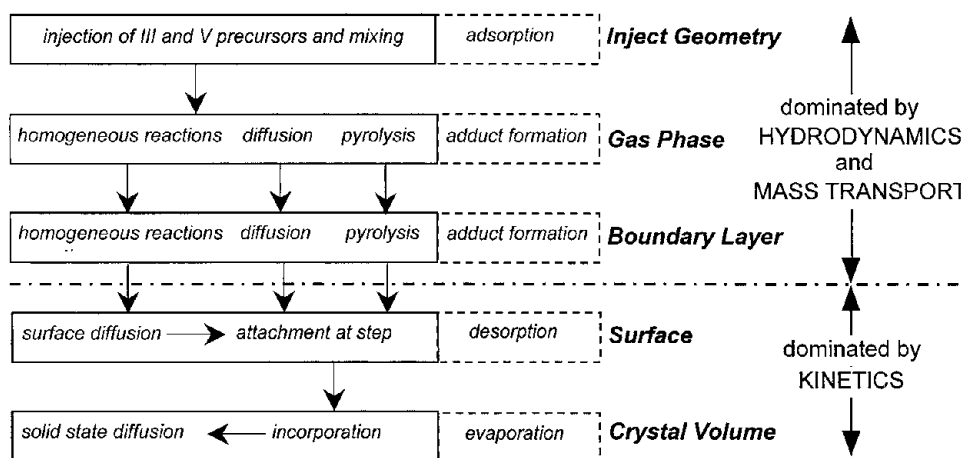


FIGURE 1 Schematic illustration of the primary reactions involved in MOCVD growth. The processes that result in reduced growth rates are in dashed boxes. Homogeneous reactions occur in the vapor phase and heterogeneous reactions occur at the interface between the vapor phase and solid phase. For some precursors, adduct formation in the vapor phase can greatly reduce the growth efficiency of the process.

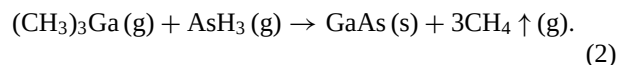
1. Fundamental MOCVD Reactions

The detailed chemical reactions occurring under “standard” MOCVD conditions have only recently begun to be understood. A schematic diagram of the important features of the MOCVD process occurring in various “regions” in the MOCVD process is shown in Fig. 1. The specific reaction kinetics and detailed thermodynamics are strong functions of the precursors and substrate employed, as well as the growth pressure, temperature, carrier gas, and reactor geometry. The hydrodynamic characteristics of the reactor chamber can also play a significant role in the outcome of growth experiments. These complications have contributed to the general lack of a detailed understanding of the MOCVD process. However, it is generally accepted that the net chemical reactions for the growth of III–V binary compounds by MOCVD are pyrolysis-driven reactions of the form:



where M is a Column III metal atom, e.g., Ga, Al, or In; R is an organic radical, typically CH_3 or C_2H_5 ; and E is a Column V atom e.g., As, P, or N. The reactions of the type described in Reaction (1) are generally called “Lewis acid–Lewis base” reactions. The Lewis acid (electron pair acceptor) in this case is the hydride and the Lewis base (electron pair donor) is the metal atom in the metal alkyl. While this greatly simplified net reaction ignores intermediate reactions or “addition compound” formation that might occur, it provides a basic framework that can be used to describe the more complicated cases where more than one organometallic or hydride are involved, e.g., for the growth of quaternary compound semiconductors. For

example, typical “generic” net reaction employed for the MOCVD growth of GaAs is given below in Eq. (2). While this reaction seems to be composed of very simple pyrolysis reactions, more complete reaction models have been developed that include more than 39 individual intermediate reactions and byproducts.



As noted above, the MOCVD growth of III–V semiconductor films can be complicated by homogeneous reactions in the gas phase, precursor-dependent activation energies and pyrolysis efficiencies, and surface kinetics. With the advent of advanced computer modeling codes and the experimental verification of the general predictions of these models, it has recently become possible to use the results of computational fluid dynamics techniques to determine the most favorable operating regime for some reactor systems. However, for the study of specific materials and device parameters, the crystal grower is required to explore the growth parameter space peculiar to the specific reactor employed in order to determine the optimum conditions for the growth of epitaxial thin films. This is especially important for the commonly employed large-scale production reactors having growth chambers with both a vertical geometry (the “rotating-disk reactor,” or RDR), and a horizontal geometry (the “Planetary Reactor”) as described below.

2. Homogeneous and Heterogeneous Reactions

The precursor reactions that are present under most commonly used MOCVD growth conditions generally consist of chemical processes that occur in both the gas phase

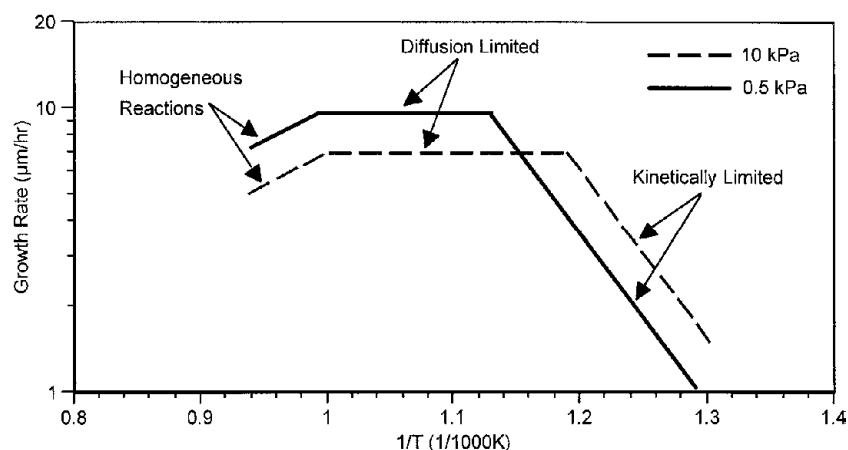


FIGURE 2 Temperature and pressure dependence of the growth rate for a typical MOCVD process. The regime where the growth is kinetically limited occurs at lower temperatures. At intermediate temperatures, the growth is limited by vapor-phase diffusion of precursors (typically the metalorganic) through the “boundary layer” near the growing surface. At higher temperatures, the growth can be affected by homogeneous reactions and deposition on the chamber walls.

(i.e., the homogeneous reactions) and at the semiconductor substrate surface (i.e., the heterogeneous reactions). Depending upon the specific precursors and the surface, the growth parameters, and the reactor geometry, either the homogeneous or the heterogeneous reactions will usually dominate the process. In many cases, the kinetic, thermodynamic, and hydrodynamic processes can be simply modeled, and basic assumptions can lead to useful predictions regarding growth rates and which of the many chemical reactions is likely to be the rate-limiting step. The dominance of homogeneous or heterogeneous reactions is usually a strong function of growth temperature, as shown schematically below in Fig. 2 for a “typical” MOCVD growth process. In general, growth at low temperatures ($T_g < 500^\circ\text{C}$) is kinetically limited. In the “middle range” ($550^\circ\text{C} \leq T_g \leq 750^\circ\text{C}$), the growth is usually diffusion-rate limited by diffusion of the organometallic precursor through the “boundary layer.” The “boundary layer” is the region in the gas phase near the surface of the substrate where the gas velocity decreases from the more or less constant “bulk value” in the growth chamber to essentially zero at the substrate. Growth at high temperatures ($T_g > 800^\circ\text{C}$) is often limited by homogeneous reactions and parasitic deposition on the reactor walls. These “break-point temperatures” are strongly a function of the material being grown. The examples of Fig. 2 are typical for the growth of GaAs.

A typical pressure dependence for MOCVD growth of GaAs is shown schematically in Fig. 3. For very low total pressures ($P_{\text{tot}} < 1$ kPa), the growth is entirely kinetically controlled, even at relatively high temperatures, resulting in a zero slope in the R_g vs P curve. For pressures

above 1 kPa, the growth rate is primarily controlled by diffusion through the thin boundary layer above the substrate surface, resulting in a $-1/2$ slope in the $\log R_g$ vs $\log P_{\text{tot}}$ curve. Growth in the pressure regime $P_{\text{tot}} < 1$ kPa is usually referred to as ultralow pressure MOCVD. At even lower pressures ($P_{\text{tot}} < 10$ Pa), the process is called ultrahigh vacuum (UHV) MOCVD. Growth at pressures $P_{\text{tot}} > 10$ kPa occur in a “viscous-flow” regime, whereas growth in the range $P_{\text{tot}} < 10$ Pa occurs in the “molecular-flow” mode, and is sometimes referred to as “metalorganic molecular-beam epitaxy” (MOMBE) or “chemical-beam epitaxy” (CBE). Such low pressures are required so that molecules can traverse the space between the source

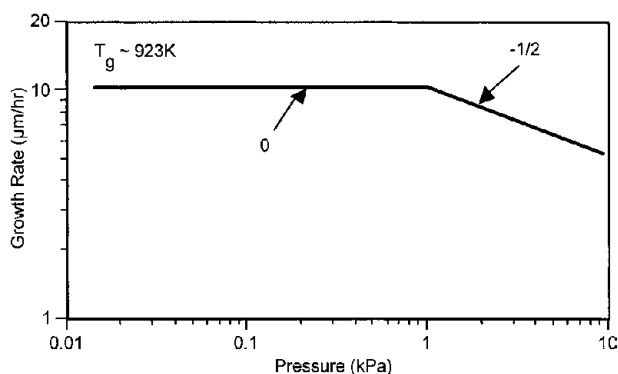


FIGURE 3 Pressure dependence of the growth rate for a typical MOCVD growth process. The growth rate is independent of pressure for the low-pressure regime ($P_{\text{tot}} < 1$ kPa) where heterogeneous reactions and surface kinetics controls the growth; at higher pressures ($P_{\text{tot}} > 1$ kPa), diffusion through the “boundary layer” is the rate-limiting step for the growth of epitaxial films.

“injector” or point of origin and the substrate surface without interacting with any other molecules (especially impurities like O_2 or CO_2).

3. Precursor Selection

The metalorganic precursor compounds that have been most commonly used to grow thin films of semiconductors and related materials are listed below in Table I, along with the currently available vapor pressure data. These precursors are typically pyrophoric liquids or high-vapor-pressure solids. The simple metal alkyls (methyl and ethyl derivatives) are the most often employed for the growth of III–V compound semiconductors since they have reasonably high vapor pressures and can be readily delivered using a H_2 carrier gas and precursor source temperatures conveniently near room temperature.

These compounds are synthesized, purified, and loaded under well-controlled conditions into specially designed and prepared all-welded stainless-steel vessels. The metalorganic precursors are transported by passing a controlled flow of the carrier gas through the precursor storage vessel and transporting the resulting vapor-phase mixture into a gas mixing system, commonly referred to as the “injection manifold” that is, in turn, connected to a mixing region at the inlet to the reaction chamber. The various precursor gases are again mixed with a high volume of the carrier gas and enter the “input zone” of the reaction chamber. The gas mixture passes over the heated substrate and thermally driven chemical reactions occur, both in the gas phase (i.e., homogeneous reactions) and at the vapor–solid interface (i.e., heterogeneous reactions). Often the homogeneous reactions can lead to the formation of undesirable intermediate compounds (e.g., adducts) formed between the Column III and Column V precursors. These adducts typically have extremely low vapor pressures and do not react to produce epitaxial materials, resulting in a reduction in the effective molar flow of useable precursors and a corresponding reduction of the growth rate.

4. General Description of MOCVD Growth Systems

The early vertical-geometry MOCVD reactors operated at atmospheric pressure (760 Torr or 10^5 Pa) and consisted of a quartz chamber with a slowly rotating (~ 5 – 20 rpm) SiC-coated graphite “susceptor” upon which the substrate was placed. Atmospheric-pressure horizontal growth systems employing circular cross section quartz chambers were also used. In most cases, induction heating of the graphite susceptor was provided by an RF generator.

Today, most multiple-wafer MOCVD systems are operated at sub-atmospheric pressure, in the 20–300 Torr (2.6–40 kPa) range to improve the uniformity of the epi-

taxial films. These low-pressure MOCVD systems operate under controlled pressures using chemical-series vacuum pump systems (now “dry” oil-free pumps are often used) to pull the reactants through the chamber at high gas velocities, thus reducing the boundary layer thickness and reducing the gas switching time in the chamber. The first work on low-pressure MOCVD (LP-MOCVD) was reported by J. P. Duchemin *et al.* (Thompson CSF, France) in 1979, who reported the growth of InP using triethylindium (TEIn) and PH_3 , and GaAs using triethylgallium (TEGa) and AsH_3 , at ~ 100 Torr (~ 13 kPa) in a horizontal reactor. This same group also reported the growth of InGaAsP alloys lattice-matched to InP at low pressure using TEGa, TEIn, PH_3 , and AsH_3 . Using LP-MOCVD, they also grew the first InGaAsP/InP injection lasers produced by MOCVD.

The MOCVD reactors that are in primary use today are generally of one of two types: (1) a cylindrical cold-wall stainless-steel reactor chamber using high-speed rotation ($R_{rot} \geq 500$ rpm) of a resistance-heated molybdenum or graphite wafer carrier inside the chamber. Most current-generation vertical-geometry reactors using high-speed rotation to produce a uniform temperature profile, a thin boundary layer, and well-developed laminar-flow gas streamlines. These chambers are based on the classical RDR (see Fig. 4). Or (2) a rectangular cross-section cold-wall quartz-walled chamber employing RF or lamp heating of a graphite susceptor that, in addition to the rotation of the main “wafer platter,” employs “gas-foil” rotation of individual wafers ($R_{rot} \sim 1$ – 3 rpm) to improve the uniformity of the growth (see Fig. 5). Advanced horizontal-geometry reactors of this type are also available commercially. The large chambers of this

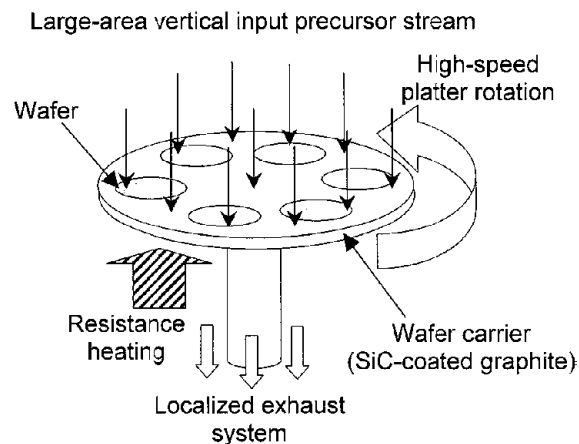


FIGURE 4 Schematic diagram of a typical large-scale high-speed vertical rotating-disk MOCVD reactor chamber including a simplified view of gas flow in a vertical RDR. The inlet gas stream contains the precursor flows and the main carrier gas flow. Typically, the Column V and Column III sources are kept separate until a few inches above the heated susceptor.

TABLE I List of Chemical Formulas and Physical Properties of Metal Alkyls

Element	Metalorganic source	Chemical formula	Vapor pressure $\log P = B - A/T$ (Torr)
Aluminum	Diisobutylaluminum hydride	$(C_4H_9)_2AlH$	—
	Dimethylaluminum hydride	$(CH_3)_2AlH$	$B = 8.92, A = 2575$
	Ethyltrimethylamine alane	$(CH_3)_2C_2H_5NAlH_3$	—
	Triethylaluminum	$(C_2H_5)_3Al$	$B = 8.999, A = 2361.2$
	Triisobutylaluminum	$(C_4H_9)_3Al$	$B = 7.121, A = 1710.3$
	Trimethylaluminum	$(CH_3)_3Al$	$B = 8.224, A = 2134.83$
	Trimethylamine alane	$(CH_3)_3NAlH_3$	—
	Tertertiarybutylaluminum	—	$P = 2 \text{ Torr @ } 300 \text{ K}$
Antimony	Triethylantimony	$(C_2H_5)_3Sb$	$B = 7.904, A = 2183$
	Triisopropylantimony	$(C_3H_7)_3Sb$	$B = 9.268, A = 2881$
	Trimethylantimony	$(CH_3)_3Sb$	$B = 7.7068, A = 1697$
	Tris-dimethylaminoantimony	—	—
Arsenic	Tertiarybutylarsine	$(C_4H_9)AsH_2$	$B = 7.5, A = 1562.3$
	Tetraethyl biarsine	$(C_2H_5)_4As_2$	—
	Triethylarsenic	$(C_2H_5)_3As$	$B = 8.23, A = 2180$
	Trimethylarsenic	$(CH_3)_3As$	$B = 7.405, A = 1480$
	Phenylarsine	—	—
	Tris-dimethylaminoarsenic	—	—
Barium	Bariumhexafluoroacetylacetonate	$(CF_3COCHCOCF_3)_2Ba$	—
Beryllium	Diethylberyllium	$(C_2H_5)_2Be$	$B = 7.59, A = 2200$
Bismuth	Trimethylbismuth	$(CH_3)_3Bi$	$B = 7.628, A = 1816$
Boron	Triethylboron	$(C_2H_5)_3B$	$B = 7.413, A = 1544.2$
Cadmium	Dimethylcadmium	$(CH_3)_2Cd$	$B = 7.764, A = 1850$
	Diethylcadmium	$(C_2H_5)_2Cd$	—
Carbon	Carbon tetrabromide	CBr_4	$B = 7.7774, A = 2346.14$
	Carbon tetrachloride	CCl_4	—
Cobalt	Tricarbonylnitrosylcobalt	$Co(NO)(CO)_3$	—
Copper	Copper hexafluoroacetylacetonate	$(CF_3COCHCOCF_3)_2Cu$	—
	Cyclopentadienylcopper triethylphosphine	$(C_5H_5)(CuP)(C_2H_5)_3$	—
	Copper (hexafluoroacetylacetonate)(1.5-cyclooctadiene)	$Cu(CF_3COCHCOCF_3)(C_8H_{12})$	—
Erbium	Tris(methylcyclopentadienyl) erbium	$(CH_3C_5H_4)_3Er$	—
Gallium	Diethylgallium chloride	$(C_2H_5)_2GaCl$	—
	Triethylgallium	$(C_2H_5)_3Ga$	$B = 8.224, A = 2222$
	Trimethylgallium	$(CH_3)_3Ga$	$B = 8.07, A = 1703$
	Triisopropylgallium	$(C_3H_7)_3Ga$	—
	Triisobutylgallium	$(C_4H_9)_3Ga$	$B = 4.769, A = 1718$
	Tetramethylgermanium	$(CH_3)_4Ge$	$B = 7.879, A = 1571$
Indium	Ethyltrimethylindium	$(CH_3)_2(C_2H_5)In$	—
	Triethylindium	$(C_2H_5)_3In$	$B = 8.93, A = 2815$
	Trimethylindium	$(CH_3)_3In$	$B = 10.52, A = 3014$
	Trimethylindium-trimethylphosphine Adduct	$(CH_3)_3In-P(CH_3)_3$	$B = 6.9534, A = 1573$
	—	—	—
Iodine	Methyliodide	CH_3I	$B = 7.684, A = 1514.5$
	Ethyliodide	C_2H_5I	$B = 7.877, A = 1715$
Iron	Bis(cyclopentadienyl) iron	$(C_5H_5)_2Fe$	$B = 10.27, A = 3680$
	Pentacarbonyliron	$Fe(CO)_5$	$B = 8.514, A = 2105$
Lead	Tetraethyllead	$(C_2H_5)_4Pb$	$B = 9.0983, A = 2824$
Magnesium	Bis(cyclopentadienyl) magnesium	$(C_5H_5)_2Mg$	$B = 25.14, A = 4198$
	Bis(methylcyclopentadienyl) magnesium	$(CH_3C_5H_4)_2Mg$	$B = 7.302, A = 2358$

continues

TABLE I (continued)

Element	Metalorganic source	Chemical formula	Vapor Pressure $\log P = B - A/T$ (Torr)
Manganese	Tricarbonyl(methylcyclopentadienyl) Manganese	$(\text{CO})_3(\text{CH}_3\text{C}_5\text{H}_4)\text{Mn}$	—
Mercury	Dimethylmercury	$(\text{CH}_3)_2\text{Hg}$	$B = 7.575, A = 1750$
Neodimium	Tris(methylcyclopentadienyl) neodimium	$(\text{CH}_3\text{C}_5\text{H}_4)\text{Nd}$	—
Niobium	Niobium ethoxide	$\text{Nb}(\text{C}_2\text{H}_5\text{O})_5$	—
Nitrogen	Tertiary-butylamine	$(\text{CH}_3)_3\text{CNH}_2$	$B = 7.61, A = 1509.8$
	Phenylhydrazine	$\text{C}_6\text{H}_5\text{NHNH}_2$	$B = 8.749, A = 3014$
	Dimethylhydrazine	$(\text{CH}_3)_2\text{NHNH}_2$	—
Phosphorus	Diethylphosphine	$(\text{C}_2\text{H}_5)_2\text{PH}$	$B = 7.6452, A = 1699$
	Mono- <i>t</i> -butylphosphine	$(\text{C}_4\text{H}_9)\text{PH}_2$	$B = 7.586, A = 1539$
	Tertiarybutylphosphine	$(\text{C}_4\text{H}_9)\text{PH}_2$	—
	Tris-dimethylaminophosphorous	—	—
Selenium	Diethylselenide	$(\text{C}_2\text{H}_5)_2\text{Se}$	$B = 7.905, A = 1924$
	Diisopropylselenide	$(\text{C}_3\text{H}_7)_2\text{Se}$	$B = 7.558, A = 1946$
	Dimethylselenide	$(\text{CH}_3)_2\text{Se}$	$P(\text{mmHg}) = (7.98 \pm 0.25) - (1678 \pm 78)/T \text{ (K)}$
Silicon	Silicon tetrachloride	SiCl_4	—
	Tetraethoxysilane (TEOS)	$(\text{C}_2\text{H}_5\text{O})_4\text{Si}$	$B = 6.88, A = 1770$
	Silicon tetrabromide	SiBr_4	—
Sulfur	Diethylsulfide	$(\text{C}_2\text{H}_5)_2\text{S}$	$B = 8.184, A = 1907$
	Propylene sulfide	$(\text{C}_3\text{H}_6)\text{S}$	$B = 6.91, A = 1405$
	Diisopropylsulfide	$(\text{C}_3\text{H}_7)_2\text{S}$	$B = 7.558, A = 1946$
Tantalum	Tantalum ethoxide	$\text{Ta}(\text{C}_2\text{H}_5\text{O})_5$	—
Tellurium	Diallyltelluride	$(\text{C}_3\text{H}_5)_2\text{Te}$	$B = 7.308, A = 2125$
	Diethyltelluride	$(\text{C}_2\text{H}_5)_2\text{Te}$	$B = 7.99, A = 2093$
	Diisopropyltelluride	$(\text{C}_3\text{H}_7)_2\text{Te}$	$B = 8.125, A = 2250$
	Dimethylditelluride	$(\text{CH}_3)_2\text{Te}_2$	$B = 6.94, A = 2200$
	Dimethyltelluride	$(\text{CH}_3)_2\text{Te}$	$B = 7.97, A = 1865$
	Di- <i>t</i> -butyltelluride	$(\text{C}_4\text{H}_9)_2\text{Te}$	$B = 4.727, A = 1323$
	Methylallyltelluride	$(\text{CH}_3)(\text{C}_3\text{H}_5)\text{Te}$	$B = 8.146, A = 2196$
Thallium	Cyclopentadienylthallium	$(\text{C}_5\text{H}_5)\text{Tl}$	$P(\text{KPa}) = 8.60 \pm 0.5 - (3706 \pm 150)/T$
Tin	Tetraethyltin	$(\text{C}_2\text{H}_5)_4\text{Sn}$	$B = 8.9047, A = 2739$
	Tetramethyltin	$(\text{CH}_3)_4\text{Sn}$	$B = 7.445, A = 1620$
Vanadium	Vanadium triethoxide oxide	$\text{VO}(\text{C}_2\text{H}_5)_3$	—
Yttrium	Tris(methylcyclopentadienyl) yttrium	$(\text{CH}_3\text{C}_5\text{H}_4)\text{Y}$	$B = 20.45, A = 6628$
Zinc	Diethylzinc	$(\text{C}_2\text{H}_5)_2\text{Zn}$	$B = 8.28, A = 2109$
	Dimethylzinc	$(\text{CH}_3)_2\text{Zn}$	$B = 7.802, A = 1560$

design employ stainless-steel chambers that are cylindrical in shape and employ graphite wafer carriers that have a specially designed “counterrotation” planetary geometry with the individual wafers rotating in the opposite direction from the main wafer carrier. These wafers are mounted on gas-bearing-supported wafer carriers and are levitated slightly above the main wafer carrier as well as rotated by the “supporting” gas stream. Also in use in a variety of manufacturing facilities, particularly in Japan,

are “custom-designed” proprietary multiple-wafer reactor chambers employing a “barrel reactor” design. Recently, commercial MOCVD reactors of both vertical and horizontal types have become available with capacities of up to 5×6.0 , 12×4.0 , 30×3.0 , or 48×2.0 in. diameter wafers (or more) per run. Recently a horizontal Planetary reactor was announced with capacity for 95×2.0 , 25×4.0 or 5×10.0 in. diameter wafers. Some custom MOCVD reactor systems are even larger in capacity.

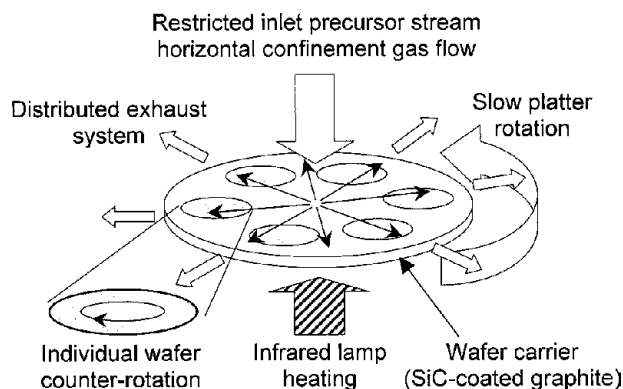


FIGURE 5 Schematic diagram of a typical large-scale horizontal "gas foil" Planetary MOCVD reactor chamber. The precursor gases are injected in the center of the rotating wafer carrier and the gas flows horizontally over the individually rotating wafers.

Commercial state-of-the-art RDR MOCVD reactors typically employ stainless-steel growth chambers that are UHV compatible and are normally fitted with a stainless-steel load-lock chamber through which wafers are loaded into the growth region using a pneumatically controlled wafer transfer arm. This greatly reduces the exposure of the growth chamber to ambient O_2 and H_2O vapor. In the horizontal MOCVD systems, this is often accomplished by enclosing the reactor chamber entry port in a glove box containing a dry N_2 ambient. Advanced MOCVD growth systems employ full computer control of the flows, pressures, temperatures, times, and valve sequences associated with the growth process. New system designs are appearing that are fully compatible with the semiconductor industry standard robotic interface. The external view of the growth chamber of a current-generation vertical RDR reactor is shown in Fig. 6 and the interior of the growth chamber of a current-generation horizontal gas-foil rotation Planetary reactor chamber is shown in Fig. 7.

II. PROPERTIES OF COMMON METALORGANICS AND HYDRIDES USED FOR MOCVD

The metal alkyls commonly used as precursors for the MOCVD growth of III-Vs can, in principle, be made by very simple halogen-containing reagent reactions. A basic example of this process is described by the following reactions for the formation of TEGa and TMGa:

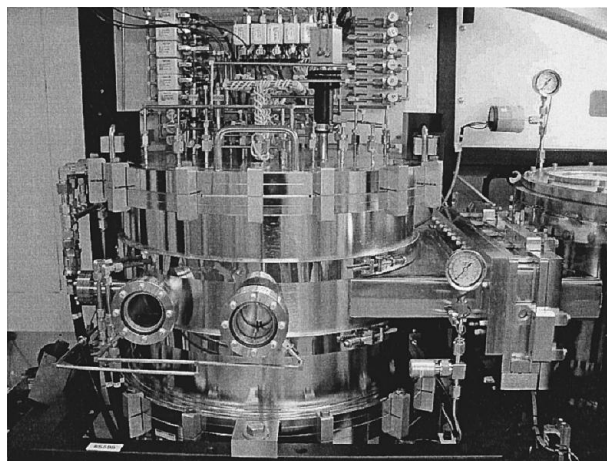
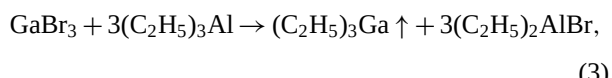


FIGURE 6 Photograph of the growth chamber of a large-scale commercial RDR MOCVD system. The gas injection manifold is shown in the rear behind the stainless steel growth chamber. The large wafer carriers are loaded into the growth chamber through the rectangular port on the right side of the chamber. The robotic interface for the robotic computer-controlled platter handling system is shown on the left. (Photograph of EMCORE Model Enterprise E450, courtesy of EMCORE Corporation.)

Many of the synthetic routes used in the early days of MOCVD involve reactions with chemicals that can subsequently provide impurity atoms in the product. For example, the above Reaction (3) can leave the TEGa with a small amount of TEAl and Reaction (4) can produce Zn-contaminated TMGa. Note that many metal alkyls are

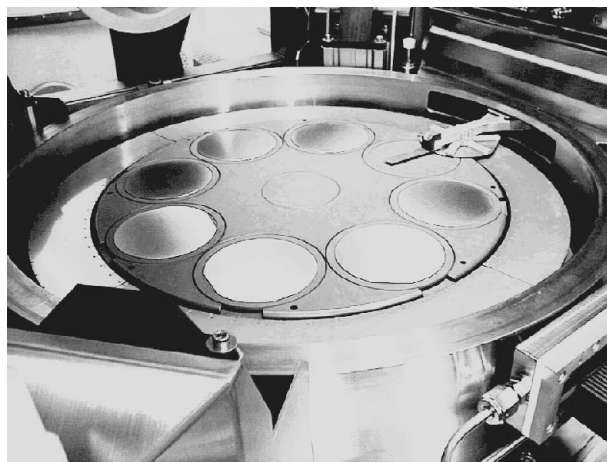


FIGURE 7 Photograph of a large-scale commercial horizontal Planetary MOCVD system. This reactor has a capacity for five planetary wafer carriers with 6-in. diameter wafer capacity. This particular example is fitted with eight gas-foil wafer carriers capable of holding one 4-in. diameter wafer and has a robotic interface that loads individual wafers on the wafer carriers. (Photograph of AIXTRON Model AIX2400/2600G3, courtesy of AIXTRON Corporation.)

prepared by reactions with Al-based metalorganics, as in example Reaction (3) above. Metal alkyls containing Al can be prepared through the use of an organo-lithium compound and aluminum trichloride, e.g.,



Such organo-aluminum compounds are used in large quantities in the chemical industry as catalysts in the manufacture of plastics, polyethylene, etc., and low-purity versions of these metal alkyls are manufactured in industrial plants in “railroad car” quantities (thousands of kilograms per month). They are also used in the manufacture of pharmaceuticals, flavor agents, and fragrances.

Simple fractional distillation processes for purification of metalorganics can be employed to remove some of these impurities, but this is a very inefficient approach. A dramatic improvement in the yield of many high-purity metal alkyl compounds resulted from the development of the “adduct-purification” scheme for the purification of metal alkyls, which was commercially developed by A. C. Jones and coworkers. This process uses the strong tendency of many metal alkyls to form stable adduct compounds with other reactants, thus making a difficult problem that is encountered in the epitaxial growth arena into an useful advantage in the synthetic arena. Actual synthetic and purification routes employed in the manufacture of metal alkyls are proprietary. It is a challenge to develop an optimized synthetic process that has the required purity, efficiency, volume, reproducibility, and yield.

For any crystal growth process, an extremely important consideration in the growth of high-quality epitaxial device structures is the purity of the sources. This is especially true for MOCVD since the organometallic precursors are extremely reactive and are thus difficult to purify. In addition, the hydrides are toxic and, in the case of PH_3 , also can react with air to form hazardous materials. Owing to these difficulties, it is only recently that techniques have been developed to *directly* measure the impurities in the metal alkyls with the necessary sensitivities in the range of parts per billion (ppb) by weight. This level of detection is required because “unintentional” impurity concentrations in the solid films grown with these sources directly influence the electronic properties of the epitaxial film. For example, a GaAs epitaxial layer having an atomic density of $\sim 2 \times 10^{22}$ atoms/cm³ and a total impurity concentration of 10 ppb of one specific element would have an unintentional concentration of “unwanted” atoms of nearly 2×10^{14} cm⁻³. In many cases, these unintentional impurities are present in organometallic sources in much higher concentrations, in the range of 2–5 ppm by weight. This results in a much higher concentration of unintentional impurities being incorporated into the semi-

conductor film. Some of the impurities in the hydrides also contribute to this problem, e.g., water vapor in the hydrides may enhance the incorporation rates of certain impurities in the metal alkyls, particularly O.

Besides the Column III metalorganics, several “alternate alkyl-containing Column V precursors” have been developed to replace the hazardous As and P hydrides, arsine (AsH_3) and phosphine (PH_3). The most practically successful of these are tertiarybutylarsine ($\text{C}_4\text{H}_9\text{AsH}_2$, TBAs) and tertiarybutylphosphine ($\text{C}_4\text{H}_9\text{PH}_2$, TBP). While these molecules have only one of the parent hydride’s H atoms replaced with a butyl group (C_4H_9), they are considerably safer to use because of the much lower vapor pressure of these liquid sources compared to the higher vapor pressures of the pure hydrides. Furthermore, the TB compounds decompose more readily at lower temperatures than the pure hydrides, making lower V/III ratios more practical, resulting in a smaller usage rate for the toxic chemicals during growth. However, these sources initially were not as pure as the AsH_3 and PH_3 parents, and they were not readily accepted in production. While currently there are large-scale users of TBAs and TBP (particularly in Japan), these precursors have not gained wide acceptance and, as a result, the cost per gram is still quite high.

In addition to these Columns III and V “primary precursors,” vapor-phase sources of dopant atoms—e.g., Zn and Mg from Column II; carbon from Column IV for *p*-type doping; and S, Si, Se, Te from Column V—for *n*-type doping are required for the growth of epitaxial device structures. In most practical applications, these dopants can be readily obtained from the corresponding precursors listed in Table I. Of particular note is the metal alkyl source for Mg, (bis)cyclopentadienylmagnesium [$(\text{C}_2\text{H}_5)_2\text{Mg}$, Cp_2Mg] (a solid source) that is commonly used to provide Mg acceptor atoms to make *p*-type wide-bandgap III–V materials (e.g., materials in the InAlGaP and InAlGaIn systems).

The availability of inductively coupled plasma mass spectrometry (ICPMS) has provided a method of detection of many impurities at very low concentrations directly in the organometallic compound itself. ICP mass spectrometry is a relatively recently developed chemical analysis technique that is useful in the detection of trace element concentrations in a liquid or solid matrix. ICPMS can measure the presence of almost all elements simultaneously, thus giving a detailed, semiquantitative picture of the impurity distribution in the sample. This technique has sensitivities for many elements in the parts-per-billion to parts-per-trillion range. It has the advantage that it is extremely sensitive and can analyze small samples (10 ml or less) of organometallics directly. The ICPMS technique employs a plasma to dissociate the material to be characterized into

TABLE II Typical Sensitivity of ICPMS for Various Metal Elements in Metal Alkyls^a

Element/ sensitivity (ppm)	Element/ sensitivity (ppm)	Element/ sensitivity (ppm)	Element/ sensitivity (ppm)
Ag < 0.4	Cr < 0.4	Mn < 0.03	Se < 1.0
Al < 0.5	Cu < 0.05	Mo < 0.5	Si < 0.03
As < 0.5	Fe < 0.1	Na < 0.5	Sn < 0.5
Au < 0.5	Ga < 0.5	Nb < 0.5	Sr < 0.1
B < 0.4	Ge < 0.5	Ni < 0.5	Tb < 0.5
Ba < 0.1	Hg < 0.5	P < 0.5	Ti < 0.2
Be < 0.02	In < 0.5	Pb < 1.0	U < 1.5
Bi < 0.1	K < 1.0	Pd < 0.5	V < 0.5
Ca < 0.02	La < 0.4	Pt < 0.5	W < 0.5
Cd < 0.02	Li < 0.4	Rh < 0.5	Y < 0.02
Co < 0.4	Mg < 0.02	Sb < 1.0	Zn < 0.2

^a Data from Air Products and Chemicals, Allentown, PA, United States.

ionized fragments that are then analyzed by a sensitive mass spectrometer (typically a magnetic sector instrument). At present, many manufacturers of electronic-grade organometallic compounds employ ICPMS to routinely analyze each batch of precursors. This has greatly reduced the variability of metal alkyl sources that are manufactured using the “same” process and equipment. Prior to the use of ICPMS, the only useful way of testing the purity of “electronic grade” organometallics was the “use test”—grow an epitaxial film using a “standard” growth run recipe and analyze the resulting film for impurities. In most cases, this involved using low-temperature photoluminescence, variable-temperature Hall-effect mobility analysis, secondary-ion mass spectrometry (SIMS), or photothermal ionization spectroscopy. All of these techniques are costly and time-consuming. In many cases, the sensitivity is inadequate to indicate the exact chemical composition of the impurities. Furthermore, the impurity concentrations can depend upon the growth conditions and the other sources used in the growth, e.g., the hydride group V sources.

Recently, many of the commonly used precursors, e.g., TEGa, TMGa, TMIIn, and TMAI, have become available in special high-purity forms from a variety of vendors. An especially important consideration for the growth of many high-quality semiconductor materials is the reduction of the oxygen-containing species in these precursors, e.g., unwanted residual alkoxide compounds. “Low-oxygen” sources have now been developed, particularly, TMAI sources. In recent work, it has been shown that the use of low-oxygen TMAI leads to an increase in the PL intensity for AlGaAs layers by a factor of 3–10 over the same alloy layers grown using “normal” grades of TMAI. Low-oxygen TMGa and TMIIn are also becoming avail-

able for critical applications requiring these sources, e.g., the MOCVD growth of LEDs and injection lasers containing InAlGaP and InAlGaIn alloys.

The selection of the Column V precursor is of equal importance. High-purity AsH₃, PH₃, and NH₃ are most commonly used and are now available from various vendors. These hydrides are extremely toxic and great care must be taken to handle them safely. Because the purity of the as-produced hydrides is not yet equal to the purity of H₂, point-of-use purifiers are normally used to ensure the purity required for high-performance devices. The threshold limit values (TLVs) established by the American Conference of Governmental Industrial Hygienists (ACGIH) for the “safe” exposure to these gases for an 8-hr period are 0.050 ppm for AsH₃, 0.3 ppm for PH₃, and 50 ppm for NH₃. Lethal concentrations for exposure of a few minutes are approximately AsH₃ ≥ 0.5 ppm, PH₃ ≥ 2 ppm, and NH₃ ~ 2000–3000 ppm. These values are listed in the corresponding Material Safety Data Sheets (MSDSs), copies of which are shipped with each cylinder of gas.

Other materials commonly used in gaseous form for the doping of MOCVD-grown films are the hydrides silane (SiH₄), disilane (Si₂H₂), germane (GeH₄), hydrogen selenide (H₂Se), hydrogen sulfide (H₂S), diethyltelluride (DETe), and the halogens carbon tetrachloride (CCl₄), and carbon tetrabromide (CBr₄). Typically, these dopant gases are supplied in high-pressure mixtures in hydrogen with dopant precursor concentrations in the 10–200 ppm range. All of these high-pressure gas sources are hazardous and extra precautions for the safe handling of gas cylinders and the disposal of reaction by-products must be made.

As noted above, in the past few years, there has been increasing interest in the use of “alternate Column V precursors” to replace the hazardous Column V hydride sources. Much of the recent work has been devoted to As- and P-organometallics, specifically, the monoalkyl-substituted hydrides tertiarybutylarsine (TBAs) and tertiarybutylphosphine (TBP). The growth of high-quality films of the III-As and III-P compound semiconductors using TBAs and TBP has been demonstrated. These sources are liquids near room temperature and can be supplied by bubbling a carrier gas through the storage vessel. The compounds are relatively low-vapor pressure liquids (see Table I) and thus they have inherently lower storage pressures at 300 K than the hydrides AsH₃ (220 PSIA, 1500 kPa) and PH₃ (607 PSIA, 4190 kPa), which are liquids at 300 K. The lower storage pressure of TBAs (~110 Torr, 15 kPa) and TBP (~200 Torr, 26.3 kPa) near room temperature make them safer to handle since the exposure from accidental release is likely to be greatly reduced. However, the absolute toxicities of these materials are still nearly that of the corresponding hydrides and adequate procedures for the safe handling use of these

materials must be followed. The TLVs for TBAs are TBP not yet established. However, some toxicity data on TBAs and TBP has been obtained. The lethal concentrations for which 50% of the exposed rat population dies (the LC_{50} values) for TBAs and TBP are ~ 45 and ~ 1100 ppm, respectively, whereas AsH_3 has an LC_{50} of ~ 45 ppm. Thus, these tests show that TBAs is about as toxic as AsH_3 , however, because of the lower storage pressure, the use of TBAs amounts to a significant safety advantage during storage and usage. An additional advantage of TBAs and TBP from a production viewpoint is that they provide excellent performance at relatively low input V/III ratios in the vapor phase. This offers a distinct advantage in reduced consumption of precursors as well as a reduced volume of toxic waste byproducts. While the cost per gram is still quite high for TBP and TBAs, the increased volume of production, which has occurred in recent years has led to somewhat reduced pricing. Other potential advantages of these precursors are their somewhat lower pyrolysis temperatures compared to AsH_3 and PH_3 , and possibly, the ability to purify them using various organometallic purification routes.

Another emerging alternate Column V source is the N compound unsymmetric 1,1-dimethylhydrazine $(CH_3)_2N-NH_2$ (DMHy), which can be used as a low-temperature precursor for N. The vapor pressure of DMHy at 300 K is ~ 150 Torr (19.7 kPa). Using DMHy and TMGa, films of GaN have been grown at temperatures in the range from 425 to 960°C and at V/III ratios as low as 10. These conditions are quite different from those commonly used for GaN growth using TMGa and NH_3 where temperatures $\sim 1050^\circ C$ and V/III ratios ~ 3000 –5000 are used. Another application for the use of DMHy is for the MOCVD growth of GaAsN and InGaAsN alloys. These III–V compounds are potentially useful for the realization of GaAs-based injection lasers and photodetectors working in the $1.33 \mu m < \lambda < 1.55 \mu m$ range.

III. GROWTH OF III–V COMPOUND SEMICONDUCTORS BY MOCVD

Virtually all of the III–V compound semiconductors have been grown by MOCVD, in many cases, using a variety of organometallic precursors for Column III sources. In addition, in some cases, “all organometallic” processes have been demonstrated where both the Columns III and V sources are metalorganics. A general overview of the details of these processes, as well as papers describing recent advances in the field are given in publications listed in the Bibliography. Brief summaries of the processes for various specific III–V materials are given below.

As noted above, the first high-performance heterojunction devices grown by MOCVD were AlGaAs/GaAs solar cells and injection lasers reported by Dupuis *et al.* in 1977. Since that time, MOCVD has been used to produce a variety of other important devices including light-emitting diodes, heterojunction field-effect transistors (HFETs), heterojunction bipolar transistors (HBTs), *p-i-n* photodetectors, metal–semiconductor–metal photodetectors, waveguides, light modulators, and more sophisticated integrated device structures containing multiple devices grown in one or more successive growth runs. One particularly important recent development is the expansion of the MOCVD growth of III–N materials, a process also pioneered by Manasevit *et al.* in 1971. This application of MOCVD will soon lead to the dramatic expansion of LED-based lighting products into many of the “mass-market” lighting applications, including the development of high-efficiency white-light solid-state lamps.

A. III–V Compound Semiconductors

Epitaxial films of the compound semiconductors from Columns IIIA and Column VA (also called Columns 13 and 15 according to the IUPAC labeling) of the Periodic Table are of interest for a variety of electronic and optoelectronic applications. GaAs was the first of the III–Vs to be identified as a semiconductor in about 1950. First produced in 1967, thin films of GaAs were also the first epitaxial layers of the III–V semiconductors to be grown by MOCVD. These materials can be grown in binary, ternary, quaternary, and pentanary forms. Descriptions of the MOCVD growth processes for the most commercially important III–Vs are given below.

1. GaAs and AlGaAs

As noted above, thin films of GaAs were the first epitaxial semiconductor layers grown by MOCVD. The most commonly used metal alkyl Ga sources are TMGa and TEGa, and the As precursors predominantly used are AsH_3 and TBAs. Growth temperatures are in the range $600^\circ C < T_g < 800^\circ C$. Typically, V/III ratios in the range of 50–100 are used for AsH_3 growth. Lower ratios in the $20 < V/III < 40$ are used for TBAs. Generally, higher concentrations of unintentional C acceptors are incorporated when TMGa is used compared to TEGa. This is because TMGa pyrolysis occurs by successive dissociation of CH_3 radicals, leading to C incorporation, while TEGa undergoes β -hydride elimination reactions.

In 1971, Manasevit, using TMGa, TMAI, and AsH_3 , was the first to report the growth of AlGaAs alloys by MOCVD. Since this time, AlGaAs has been grown with a variety of organometallic Column III sources, including TMGa, TEGa, TMAI, TEAl, trimethylamine, and

(TMAA), and tritertiarybutylaluminium (TTBAI). Oxygen contamination in AlGaAs films has been a continuing problem. In general, O incorporation is a function of the growth temperature, substrate orientation, and V/III ratio, with larger values of T_g , substrate misorientation, and V/III ratio resulting in lower O contamination. Typically, AlGaAs is grown at $720^\circ\text{C} < T_g < 800^\circ\text{C}$ and $V/\text{III} \geq 150$. With the advent of “low-alkoxide” grades of TMAI, *in situ* purification of AsH_3 , and the improved performance characteristics of current-generation MOCVD reactors, the O concentration (as measured by SIMS) of $\text{Al}_x\text{Ga}_{1-x}\text{As}$ ($x \sim 0.20$) films is typically below $\sim 2 \times 10^{17} \text{ cm}^{-3}$.

The first high-performance III–V devices grown by MOCVD were AlGaAs–GaAs double-heterostructure (DH) and quantum-well (QW) injection lasers. Since this early work, MOCVD has become the materials technology of choice for the large-scale growth of high-quality AlGaAs–GaAs injection lasers. For example, most of the compact-disc injection lasers, and virtually all of the high-power semiconductor lasers, are manufactured from MOCVD-grown materials.

2. InAlGaAsP/InP

The epitaxial growth of thin films of the quaternary alloys $\text{In}_x\text{Ga}_{1-x}\text{As}_y\text{P}_{1-y}$ lattice matched to InP substrates is of interest for a variety of commercially important semiconductor devices, including injection lasers and high-speed photodiodes used in high-speed long-distance optical communications systems. The MOCVD growth of these materials is normally accomplished using reactors operating at low-pressure owing to the tendency for adduct formation between TMIn and PH_3 at atmospheric pressure. The most commonly used sources are TMGa, TEGa, TMIn, AsH_3 , and PH_3 . The “alternate” precursors TBAs and TBP have also been used. In early work, Duchemin *et al.* used TEIn, TEGa, AsH_3 , and “precracked” PH_3 to grow epitaxial quaternary films on InP. Subsequently, it was discovered that the precracking was not necessary since the pyrolysis efficiency of PH_3 is greatly enhanced by surface kinetics and by the presence of TMIn. Alloy films have been grown throughout the composition range having a close lattice match to InP. In particular, InGaAsP alloys with bandgap energies corresponding to emission wavelengths of $\lambda \sim 1.2$, 1.33, and $1.55 \mu\text{m}$ have been grown by MOCVD, and are of great interest for devices for optical communications.

Another quaternary in this system is $\text{In}_{1-y}(\text{Al}_x\text{Ga}_{1-x})_y\text{As}$, which can be grown lattice matched to InP substrates. These materials are grown by MOCVD using the precursors cited above with the addition of TMAI. Recently, this quaternary has shown promise for the growth of advanced high-performance injection lasers operating

at $\lambda \sim 1.3 \mu\text{m}$. The growth of InAlGaAs/InP strained-quantum-well lasers by MOCVD has the potential to increase the performance of low-cost optical communications links, which do not require “active” temperature control, and cooling of the laser itself.

The ternaries $\text{In}_x\text{Ga}_{1-x}\text{As}$ and $\text{In}_x\text{Al}_{1-x}\text{As}$ are other important compounds, which can be grown lattice matched to InP substrates. These materials can be used to grow a variety of high-speed optoelectronic devices, including strained-quantum-well lasers, *p-i-n* photodiodes, heterojunction avalanche photodetectors, high-electron-mobility transistors (HEMTs), pseudomorphic high-electron-mobility transistors (PHEMTs), and heterojunction bipolar transistors (HBTs). Great success has been achieved in growing these device structures by MOCVD.

3. InGaAsP/GaAs

Thin films in the $\text{In}_x\text{Ga}_{1-x}\text{As}_y\text{P}_{1-y}$ quaternary system have also been grown lattice matched to GaAs substrates using MOCVD. For growth by MOCVD, the commonly used sources are again TMGa, TEGa, TMIn, AsH_3 , and PH_3 . One of the most important potential applications for these materials is to the growth of “Al-free” injection lasers operating at $\lambda \sim 0.980 \mu\text{m}$, a spectral region that is well suited to the fabrication of semiconductor lasers designed to pump solid-state lasers.

4. InAlGaP/GaAs

The $\text{In}_x(\text{Al}_y\text{Ga}_{1-y})\text{P}$ quaternary has been grown by MOCVD to fabricate a variety of visible light-emitting devices. Growth of InAlGaP films lattice matched to GaAs substrates has been accomplished by both atmospheric-pressure (AP) and low-pressure (LP) MOCVD. Currently, most work is carried out at pressures of 60–76 Torr ($\sim 10 \text{ kPa}$). The Column III and P precursors are carried into the growth zone by a high flow of H_2 carrier. Typically, TMIn, TEGa, TMAI, and PH_3 are used as precursors. In most cases, the best layer quality is obtained when a thin (~ 20 – 100 nm) “buffer layer” of GaAs is grown first. It has been found that high V/III ratios (> 400) and high growth temperatures ($T_g > 700^\circ\text{C}$) are important for the reduction of O incorporation and the activation of Mg acceptors. Si and Te are commonly used donors, usually supplied by silane (SiH_4), and DETe, respectively. At this time, MOCVD is the technology of choice for the production of InAlGaP materials for high-performance red and yellow LEDs and injection lasers emitting in the red spectral region ($\lambda \sim 630$ – 670 nm). These high-brightness LEDs have luminous efficiencies that actually exceed the output efficiency (i.e., $> 40 \text{ lumens/watt}$) of 30 W halogen lamps for the production of light. The application of MOCVD

to the production of LEDs, and the invention of new chip geometries and mounting techniques, has advanced the performance of these devices toward the realization of “the ultimate lamp.”

5. GaAsP/GaP

Alloys in the $\text{GaAs}_x\text{P}_{1-x}$ system have been grown by MOCVD using TMGa and AsH_3 and PH_3 . Both GaAs and GaP substrates have been used. While high-quality materials have been produced by MOCVD, the commercial production of these materials is still dominated by the established (and low-cost) Ga–HCl– AsH_3 vapor-phase epitaxy process.

6. Sb-Containing III–Vs

One of the III–Vs that has recently received increased attention is the growth of Sb-containing compounds by MOCVD, e.g., the materials in the InAlGaAsSb system. Typically, TMIIn, TMSb, TESb, TMAI, TMGa, and AsH_3 are used as sources. The Sb-containing materials are generally of interest for photodetectors operating in the 2–5 μm spectral region and for InAs–GaSb transistors. One problem in the growth of these materials is that they melt at relatively low temperatures. In addition, there are severe miscibility gaps in the Sb-based III–V systems. A relatively new application for Sb-containing materials is the growth of strained layers of specific InGaAsSb alloys on GaAs substrates for use in “long-wavelength” injection lasers operating at $\lambda \sim 1.33 \mu\text{m}$.

7. Materials in the InAlGaN System

Recently, great success has been achieved in the growth of III–V nitride films in the $\text{In}_y(\text{Al}_x\text{Ga}_{1-x})_{1-y}\text{N}$ quaternary system by MOCVD. The sources typically employed are TMGa, TMAI, TMIIn, and NH_3 , although TEGa, TEAl, and TBN have also been used. Since bulk GaN substrates are not yet available commercially, (0001)-oriented sapphire (or 6H-SiC) substrates are usually employed. The use of a thin ($t \sim 20 \text{ nm}$) low-temperature GaN or AlN “buffer layer” ($T_g \sim 450\text{--}500^\circ\text{C}$) is generally required to obtain high-quality heteroepitaxial growth. Growth temperatures in the range $T_g \sim 1050^\circ\text{C}$ are used for the device-quality AlGaN and GaN layers, while $T_g \sim 750\text{--}800^\circ\text{C}$ is used for InGaN alloys. Recently, heteroepitaxial films of InAlGaN films have been grown by MOCVD at temperatures $T_g \sim 800^\circ\text{C}$ in the range MOCVD is currently the only materials growth technology with the demonstrated ability to produce high-performance AlGaN–InGaN green and blue LEDs and also injection lasers operating in the $390 < \lambda < 420 \text{ nm}$ at 300 K. Recently, high-efficiency

LEDs emitting white light have been developed using nitride devices and UV phosphors. These devices have the potential to provide very high efficiency lighting and illumination. In addition, using specially grown structures, continuous-wave operation of InGaN/GaN injection lasers at 300 K for over 10,000 hr has been demonstrated by Nakamura *et al.* (Nichia Chemical Co., Japan). These devices will find application as the light source in high-density and high-capacity digital versatile disk (H-DVD) players in the near future.

8. Materials in the InGaAsN System

As mentioned previously, adding significant quantities ($>2\%$) of N to GaAs or InGaAs produces a dramatic reduction in the bandgap energy of the semiconductor. Because of extreme bandgap “bowing” in these alloy systems, the addition of a few percent of N to the GaAs or to InGaAs alloys results in a significantly smaller bandgap energy than is found for the N-free compounds. It is found that the practical limit to the incorporation of N is, however, only about 4%. High-quality thin films of these materials can be grown on GaAs substrates if the lattice mismatch created by the addition of N (or In + N) is not too great. This provides the potential for the realization of the growth on GaAs substrates of an injection laser emitting at $\sim 1.33\text{--}1.55 \mu\text{m}$. Recently, the MOCVD growth of such lasers operating continuously at 300 K has been demonstrated. This research result is not yet fully developed for commercial applications, but it could make a significant impact in optical communications systems because it is relatively low cost compared to alternate approaches and the use of GaAs substrates makes the integration of long-wavelength lasers with GaAs-based electronic circuits much more feasible.

Another application of this alloy system is for an $\sim 1 \text{ eV}$ bandgap $p\text{--}n$ junction for use in high-efficiency multiple-junction solar cells. Lattice matching to GaAs substrates can be achieved by the incorporation of appropriate concentrations of both In and N in the InGaAsN quaternary alloy films. MOCVD-grown solar cells with an $\sim 1 \text{ eV}$ bandgap energy have been demonstrated using this material.

IV. SOME REPRESENTATIVE OTHER MATERIALS GROWN BY MOCVD

A. IV–VI Semiconductor Compounds

The compounds in the $\text{Pb}_{1-x}\text{Sn}_x\text{Te}$ ternary alloy system are candidates for photodetectors in the midinfrared portion of the spectrum. Manasevit *et al.* were the first

to grow IVA–VIA compounds by MOCVD in 1975. Since this time, the MOCVD process has been used to grow thin films of many of these alloys. The precursors employed are typically tetraethyllead (TEPb), TESn, and H₂Te. However, most work in this area has been halted since these materials are somewhat unstable and other semiconductor compounds can cover the same spectral region.

B. II–VI Semiconductor Compounds

The semiconductors composed of elements in Columns IIB and VIA consist of materials covering the “wide-bandgap” region and the “narrow-bandgap” compound semiconductors. The wide-bandgap materials are those in the ZnMgSSe/ZnCdSSe systems. The wide-bandgap II–VI compounds in the Zn_xCd_{1-x}S_ySe_{1-y}/Zn_xMg_{1-x}S_ySe_{1-y} system have been grown by MOCVD. The commonly used sources are DEZn, dimethylselenium (DMSe), ditertiarybutylselenide (DTBSe), DMCD, bismethyl-cyclopentadienyl-magnesium [(MeCp)₂Mg], diethyl sulfide (DES), and H₂S. These materials are useful for visible LEDs and laser diodes. However, to date, difficulties in the MOCVD growth of high-conductivity *p*-type materials in this system has prevented the demonstration of LEDs with characteristics comparable to those fabricated from MBE-grown materials.

The most important IIB–VIA narrow-gap materials are in the HgCdZnTe quaternary system, with the ternary Hg_xCd_{1-x}Te (MCT) being the most commonly used for photodetectors in the 8–12 μm regime. The narrow-gap materials can be grown by MOCVD using elemental Hg, dimethylmercury (DMHg), diethyltelluride (DETe), methylallyltelluride (MATE), diisopropyltelluride (DIPTe), dimethylcadmium (DMCd), and dimethylzinc (DMZn). These semiconductors are all low-melting-point materials and are typically grown in the 350°C ≤ *T_g* ≤ 450°C range. Recently, the RDR multiple-wafer reactor geometry has been adapted to provide for large-area growth of uniform layers in the HgCdTe system.

C. Growth of Oxides by MOCVD

A variety of oxides have been grown by MOCVD, including the important class of high-temperature superconducting Cu oxides. Particular attention has been given to the BiSrCaCuO and YBaCuO systems. Superconducting metal oxide films grown by MOCVD have been limited in performance largely by the relatively primitive state of the novel precursors used for these materials, and by the need to develop reactor designs compatible with the low vapor pressures of these materials, and the oxidizing nature of the growth ambient.

Another important class of oxides that have been grown by MOCVD are the ferroelectrics, including PbTiO₃, BaTiO₃, PbLaZrTiO₃, and PbZr_{1-x}Ti_xO₃. This work is still in its infancy, however, promising results have been achieved. Further studies of the relationship between film properties, the mechanism of deposition, growth parameters, and the choice of precursors are necessary to discover an optimized MOCVD process for this class of important ferroelectrics thin films which will be of great use in the next generation of deep-submicron Si device design and manufacture. Another dielectric material that has been grown by MOCVD is ZnO.

D. Deposition of Metals by MOCVD

An important new application for MOCVD is the deposition of pure metal films for semiconductor integrated circuit applications. Important metals deposited by MOCVD include Al, Cu, CuAl alloys, and W films using precursors listed in Table 1. It is expected that this application area for MOCVD will expand rapidly in the next few years as the demand increases for high-density metal interconnects for Si integrated circuit technology. High-purity Al metal films have also been grown by MOCVD.

V. OTHER DEVELOPMENTS IN MOCVD

The limitations of space and the specific subject of this article have prevented me from describing many of the other important developments leading to the breadth and success of the current MOCVD technology. An important development mentioned above is the use of advanced chemical kinetics, surface kinetics, and hydrodynamics models that can provide for full three-dimensional solutions to the multifaceted boundary conditions occurring in a CVD system. The Sandia National Laboratories (USA) CVD Sciences Group, particularly M. E. Coltrin, has provided tools for the detailed analysis of MOCVD systems, and they have contributed greatly to the understanding of the large-area commercial MOCVD reactors in common use today. These chemical process models have become commercially available and are now offered by several companies. Using these models, important modifications to reactor designs have been made that greatly improve growth efficiencies, material uniformity, interface abruptness, and materials quality. This is especially important for the design and optimization of very large-scale MOCVD reactors, e.g., systems with capacities for seven 6.0 in. diameter wafers as shown in Figs. 6 and 7. In fact, these large-scale reactors are difficult and expensive to build, even in prototype form. Evaluating reactor designs through

software simulations is by far more effective and efficient (and less costly) than the old-fashioned “cut and try” method commonly used in the early “frontier” days of MOCVD reactor development.

VI. FUTURE VISION

The MOCVD process has been used for a wide variety of III–V binary, ternary, quaternary, and pentanary semiconductor films. It has also been used for the growth of oxides, superconductors, dielectrics, and the deposition of metal films, including Cu interconnects on Si integrated circuits. We can expect that the usage in all these areas will increase dramatically in the next few years. It is clear that the future development of MOCVD will continue to rely on improvements in the purity of precursors (both organometallics and hydrides). Furthermore, advances in the understanding of chemical reactions, hydrodynamics, precursor kinetics, etc., should lead to improved large-scale reactor designs capable of growing simultaneously on more than a dozen 6-in. diameter substrates using a wide range of growth pressures. Furthermore, the efficiencies of scale in the production of metal alkyls should permit the cost factor of precursors to be reduced. MOCVD reactors with kilogram quantities of metal alkyls are now common in production environments. Recently, an automatic filling system for metal alkyls has been developed that employs two 20-kg storage vessels and a dedicated piping system that will automatically fill the smaller “bubblers” that are placed on the individual MOCVD reactors. The current generation system will monitor up to eight MOCVD vessels simultaneously and automatically fill them to maintain a constant precursor molar flow rate over an extended period of time. This greatly reduces the requirements to change bubblers, resulting in markedly increased reactor “up time,” run reproducibility, and the more effective use of the metalorganic sources.

One important aspect of MOCVD (and all other CVD epitaxial growth processes, including CVD for Si) that still remains to be developed is “real-time monitoring and process control”—while some *in situ* monitoring techniques have been developed, most notably spectroscopic ellipsometry, spectrally resolved reflectivity, reflectance anisotropy spectroscopy, multibeam optical reflectance, and emissivity-corrected pyrometry. These techniques permit some useful degree of “real-time monitoring” but the missing element—the “real-time control”—is still sorely needed. One important component to this control loop is the monitoring of the gas phase and surface species. Techniques for determining gas-phase composition are well established, and in-

clude laser-induced fluorescence, differential mass spectrometry, and absorption spectroscopy. The measurement of surface species in a CVD environment is still problematical, although reflectance-difference spectroscopy has shown some promise. Additional complications arise due to the lack of spatial uniformity in the gas phase inside a reactor and near the growing surface. Real-time three-dimensional chemical mapping of the reactant species inside a CVD growth chamber is a daunting problem and one that will not yield easily using conventional techniques. Many more years of research and development are required to realize a true “process control” system for MOCVD. However, it is an area of continued activity and research results are being continually translated into commercial products.

VII. SUMMARY AND CONCLUSIONS

The growth of epitaxial films of the III–V compound semiconductors by MOCVD was patented in various forms prior to 1965 and first reported in the scientific literature in 1968. In the late 1970s, MOCVD was shown to be a viable technology for the growth of high-performance solar cells and sophisticated injection lasers. From this work, it was possible to predict that the MOCVD process would become an important element in the fabrication of a wide variety of high-performance semiconductor devices. Because of the economics and flexibility of the process, the quality of the materials produced, and the scalability of the technology, it has come to dominate the epitaxial growth of III–V semiconductors. Today, most optical memory and information recording systems, (e.g., CD-ROMs, DVD players, etc.) and optical communications systems employ QW injection lasers based upon MOCVD epitaxial films and high-performance visible LEDs rely almost exclusively on MOCVD materials technologies. In addition, today, high-performance digital cellular communications rely on the performance of MOCVD-grown heterojunction field-effect transistors and heterojunction bipolar transistors.

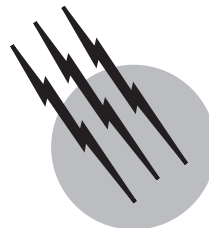
Future advances in precursor purity and manufacturing technology, real-time monitoring of chemical reactions, MOCVD reactor chamber design, computer-controlled epitaxial growth systems, detailed chemical process models, and real-time process control will lead to improved process efficiencies, reduced hazardous waste, and enhanced device reproducibility, yield, and performance. The future of MOCVD is certainly bright. We are on the frontier of a great expansion of the abilities of MOCVD to provide materials for products that improve and expand the human experience on earth, under the oceans, and in space.

SEE ALSO THE FOLLOWING ARTICLES

CHEMICAL VAPOR DEPOSITION • CRYSTAL GROWTH •
MOLECULAR BEAM EPITAXY, SEMICONDUCTORS • EX-
CITONS, SEMICONDUCTOR • LASERS, SEMICONDUCTOR •
MOLECULAR BEAM EPITAXY, SEMICONDUCTORS

BIBLIOGRAPHY

- Breiland, W. G., Coltrin, M. E., Creighton, J. R., Hou, H. Q., Moffat, H. K., and Tsao, J. V. (1999). "Organometallic Vapor Phase Epitaxy (OMVPE)." *In* "Materials Science & Engineering," vol. R24, pp. 241–274, Elsevier Science B. V., Amsterdam.
- Craford, M. G., Holonyak, N., and Kish, F. A. (2001). "In pursuit of the ultimate lamp," *Sci. Am.* **284**, 63–67.
- Dupuis, R. D. (2000). "III–V semiconductor heterojunction devices grown by metalorganic chemical vapor deposition," *IEEE J. Sel. Topics Quant. Electron.* **6**, 1040–1050.
- Jones, A. C. (1993). "Metalorganic precursors for vapour phase epitaxy," *J. Crystal Growth* **129**, 728–773.
- Kawai, H., and Onabe, K., eds. (2000). "Proceedings of the Tenth International Conference on Metalorganic Vapor Phase epitaxy, IC-MOVPEX," Elsevier Science B. V., Amsterdam.
- Stringfellow, G. B. (1999). "Organometallic Vapor Phase Epitaxy: Theory and Practice," 2nd ed., Academic Press, San Diego, CA.
- Thayer, J. S. (1998). "Organometallic Chemistry: An Overview," VCH Publishers, New York.



Pollution Prevention from Chemical Processes

Kenneth L. Mulholland

Kenneth Mulholland & Associates

Michael R. Overcash

North Carolina State University

- I. Introduction
- II. History of Pollution Prevention
- III. Waste as Pollution
- IV. How Does One Define Pollution Prevention?
- V. Drivers for Pollution Prevention
- VI. The Recipe for Success
- VII. Program Elements
- VIII. The Incentive for Pollution Prevention

- IX. Resources
- X. Engineering Evaluations of the Preferred Options
- XI. Waste Stream and Process Analyses
- XII. When Should One Do Pollution Prevention?
- XIII. Case Studies
- XIV. Conclusion

GLOSSARY

Bioaccumulative Material that accumulates in organisms, for example, lead, mercury, and DDT.

Chemical process A chemical process normally consists of a reactor section where the feed materials are reacted to the desired product(s) followed by a series of separation devices to separate the product(s) from any by-products, solvents, catalysts, etc.

Material balance Compound-by-compound listing of materials in the pipes and vessels of a process.

Persistent compound Material that does not or only slowly biodegrades, for example, PCBs and DDT.

Process flow diagram A drawing of process pipes and vessels.

I. INTRODUCTION

“POLLUTION PREVENTION” became environmental buzz words of the 1990s. No matter what one chooses to call the task or technology of reducing waste and emissions from a chemical process—pollution prevention, waste minimization, source reduction, clean technology, green manufacturing, etc.—the challenge of implementing process changes that actually reduce waste generation is often formidable. Engineers and scientists faced with developing and implementing a pollution prevention program for a business or a manufacturing site face many obstacles, technological, economic, and societal. Some of these obstacles are real, while many others are only perceived to be real.

The traditional approach to process design has been to first engineer the process and then to engineer the treatment and disposal of waste streams. However, with increasing regulatory and societal pressures to eliminate emissions to the environment, disposal and treatment costs have escalated exponentially. As a result, capital investment and operating costs for disposal and treatment have become a larger fraction of the total cost of any manufacturing process. For this reason, the *total system* must now be analyzed simultaneously (process plus treatment) to find the best economic option.

Experience in all industries teaches that processes which minimize waste generation at the source are the most economical. For existing plants, the problem is even more acute. Even so, experience has shown that waste generation in existing facilities can be significantly reduced (greater than 30% on average), while at the same time reducing operating costs and new capital investment.

In this article, we present a broad overview of the path to an effective pollution prevention program. The phases and individual steps of this proven methodology are applicable to both large-scale and small-scale problems. The focus of the methodology is on identifying pollution prevention engineering technologies and practices that will change what is happening *inside* the pipes and vessels of the manufacturing process, rather than just on simple procedural or cosmetic changes. In fact, many of the techniques and tools that support the methodology can be easily applied by chemists, process engineers, and project engineers to individual waste streams within a process or facility. For example, the methodology has been and continues to be successfully practiced inside the DuPont Company. We present a list of pollution prevention engineering technologies and practices that nicely complements the methodology and provides a useful knowledge base for quickly identifying possible process changes that reduce waste generation and emissions.

II. HISTORY OF POLLUTION PREVENTION

No single dimension of the solutions for environmental problems has captured the imagination of engineers, scientists, policy-makers, and the public like pollution prevention. In the space of two decades (1980–2000), the philosophical shift and the record of accomplishment has made pollution prevention a fundamental means for environmental management. This effort actually began during 1976–1980 when 3M Corporation initiated the 3P program and North Carolina adopted waste minimization as a state-wide priority for managing emissions from industry. By 1990, virtually all of the *Fortune* 1000 U.S. corporations had pollution prevention as the first emphasis

in describing their approach to the environment. The shift from 20–50 years of conventional pollution control to a preventative approach was dramatic because of the reversal in priorities.

The adoption of pollution prevention as a clearly differentiated approach to environmental improvement began in U.S. industry and policy during the late 1970s. While examples of improved efficiency and hence less waste had existed since the start of the Industrial Revolution, the distinct explosion of successes in pollution prevention did not occur until the mid-1980s. Figure 1 shows an approximate time line of this period.

The early creation at the 3M Corporation of money-saving innovations that reduced chemical losses to air, water, or land was widely publicized. However, propagation into other large corporations was almost nonexistent. The efforts through university research and state programs (beginning in North Carolina) to illustrate the benefits of pollution prevention, and a steady presentation of principles such as the creation of the pollution prevention hierarchy and roadmaps, extended over the early to mid-1980s. In 1986–1988, the improved information regarding chemical losses to the environment as a part of the U.S. EPA Toxic Release Inventory (TRI) Program precipitated action. A number of CEOs of large corporations challenged their companies, in a very public fashion, to reduce these chemical losses. As the autocatalytic effect spread to other

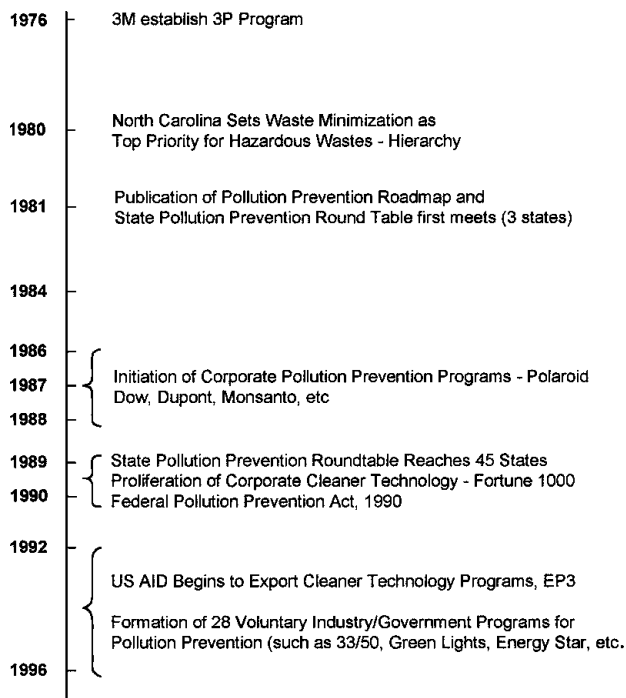


FIGURE 1 General historical sequence for growth of cleaner technology in United States.

companies and whole industry associations or sectors, the policy of priority for pollution prevention took shape in the United States. The outcome has been impressive, not necessarily uniform, but achieving a philosophical shift to cleaner manufacturing. These events are even more impressive when it is recognized that virtually all of the individual changes to manufacturing have been cost-effective (a generally held rule of a 2-year payback on capital investment).

Use of the term pollution prevention is common in the United States, but is actually one of many nearly synonymous terms, which include the following:

- Waste minimization
- Cleaner production
- Waste reduction
- Clean technology
- Source reduction
- Environmentally benign synthesis
- Environmentally conscious manufacturing
- Green chemistry
- Technology for a sustainable environment
- Sustainability
- Green engineering

Use of a particular terminology usually is linked to the forum in which the debate is occurring and hence these terms have subtle differences, but share the major emphasis on prevention. That is, all of these descriptors refer to the intuitive perspective that it is advantageous to manage chemical losses or wastes generated from the top of a hierarchy for waste minimization. In addition, there is a certain trend to reinvent terms with new government initiatives.

III. WASTE AS POLLUTION

An industrial waste is defined as an unwanted by-product or damaged, defective, or superfluous material of a manufacturing process. Most often, it has or is perceived to have no value. It may or may not be harmful or toxic if released to the environment. Pollution is any release of waste to the environment (i.e., any routine or accidental emission, effluent, spill, discharge, or disposal to the air, land, or water) that contaminates or degrades the environment.

Figure 2 depicts a typical manufacturing facility. Inputs to the facility include raw materials to produce the saleable product(s), water, air, solvents, catalysts, energy, etc. Outputs from the facility are the saleable product(s), waste energy, and gaseous, liquid, water, and solid wastes. In contrast, a manufacturing facility with an absolute minimum (but not zero) amount of waste being generated is shown in Fig. 3. Inputs to the facility include only the raw

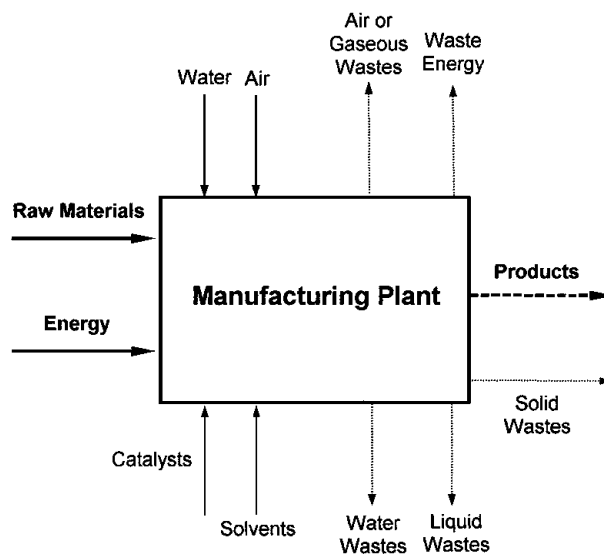


FIGURE 2 Plant with pollution.

materials to make the saleable products(s) and energy. The only significant outputs are saleable products.

IV. HOW DOES ONE DEFINE POLLUTION PREVENTION?

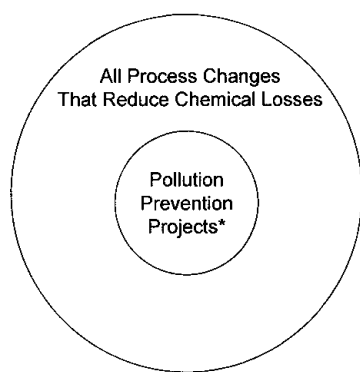
We define pollution prevention fairly broadly, in keeping with the actual practices widely utilized by industry. This definition is any cost-effective technique aimed at reducing chemical or energy-related emissions that would subsequently have to be treated. In keeping with the generally voluntary nature of U.S. pollution prevention activities, the double hurdle of technical and economic feasibility are met in a pollution prevention option (Fig. 4).

This definition manifests itself in the form of the pollution prevention hierarchy shown in Fig. 5. In this hierarchy, safe disposal forms the base of the pyramid, and minimizing the generation of waste at the source is at the peak.

The U.S. Environmental Protection Agency (EPA) definition of pollution prevention recognizes actions which encompass the upper three levels in the hierarchy: minimize generation to segregate and reuse. The U.S. EPA



FIGURE 3 Absolute minimum waste generation facility.



* Cost-Effective Changes

FIGURE 4 Context of pollution prevention within all possible process changes.

defines the hierarchy shown in Fig. 5 as environment management options. Industry defines as pollution prevention the upper five levels, from minimize generation to recover energy value in waste. The European Community, on the other hand, includes the entire hierarchy (levels 1–7) in its definition of pollution prevention, as is done in this article.

A definition of each tier in the pollution prevention hierarchy is given below:

1. *Minimize generation.* Reduce to a minimum the formation of nonsaleable by-products in chemical reaction steps and waste constituents, such as tars, fines, etc., in all chemical and physical separation steps.

2. *Minimize introduction.* Minimize the addition of materials to the process that pass through the system unreacted or that are transformed to make waste. This implies minimizing the introduction of materials that are not essential ingredients in making the final product. Examples of introducing nonessential ingredients include (1) using water as a solvent when one of the reactants, intermediates,

or products could serve the same function and (2) adding large volumes of nitrogen gas because of the use of air as an oxygen source, heat sink, diluent, or conveying gas.

3. *Segregate and reuse.* Avoid combining waste streams together without giving consideration to the impact on toxicity or the cost of treatment. For example, it may make sense to segregate a low-volume, high-toxicity wastewater stream from several high-volume, low-toxicity wastewater streams. Examine each waste stream at the source and determine which ones are candidates for reuse in the process or can be transformed or reclassified as a valuable coproduct.

4. *Recycle.* A large number of manufacturing facilities, especially chemical plants, have internal recycle streams that are considered part of the process. In this case, recycle refers to the external recycle of materials, such as polyester film and bottles, Tyvek® envelopes, paper, and spent solvents.

5. *Recover energy value in waste.* This step is a last step to attain any value from the waste. Examples include burning spent organic liquids, gaseous streams containing volatile organic compounds, and hydrogen gas for the fuel value. The reality is that often the value of energy and resources required to make the original compounds is much greater than that which can be recovered by burning the waste streams for the fuel value.

6. *Treat for discharge.* This involves lowering the toxicity, turbidity, global warming potential, pathogen content, etc., of the waste stream before discharging it to the environment. Examples include biological wastewater treatment, carbon adsorption, filtration, and chemical oxidation.

7. *Safe disposal.* Waste streams are rendered completely harmless or safe so that they do not adversely impact the environment. In this article, we define this as total conversion of waste constituents to carbon dioxide, water, and nontoxic minerals. An example is subsequent treatment of a wastewater treatment plant effluent in a private wetlands. So-called “secure landfills” would not fall within this category unless the waste is totally encapsulated in granite.

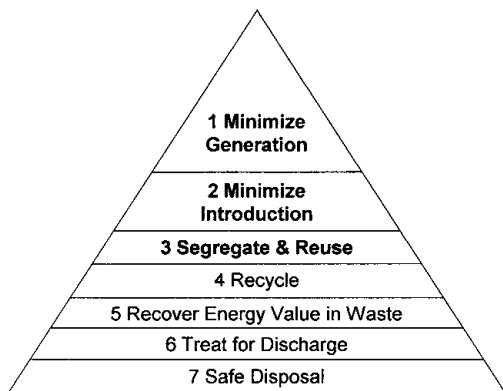


FIGURE 5 Pollution prevention hierarchy.

In this article, we will focus on the upper three tiers of the pollution prevention hierarchy; that is, minimize generation, minimize introduction, and segregate and reuse. This is where the real opportunity exists for reducing waste and emissions while also improving the business bottom line.

V. DRIVERS FOR POLLUTION PREVENTION

Since the early 1960s, the number of federal environmental laws and regulations has been increasing at a rate three

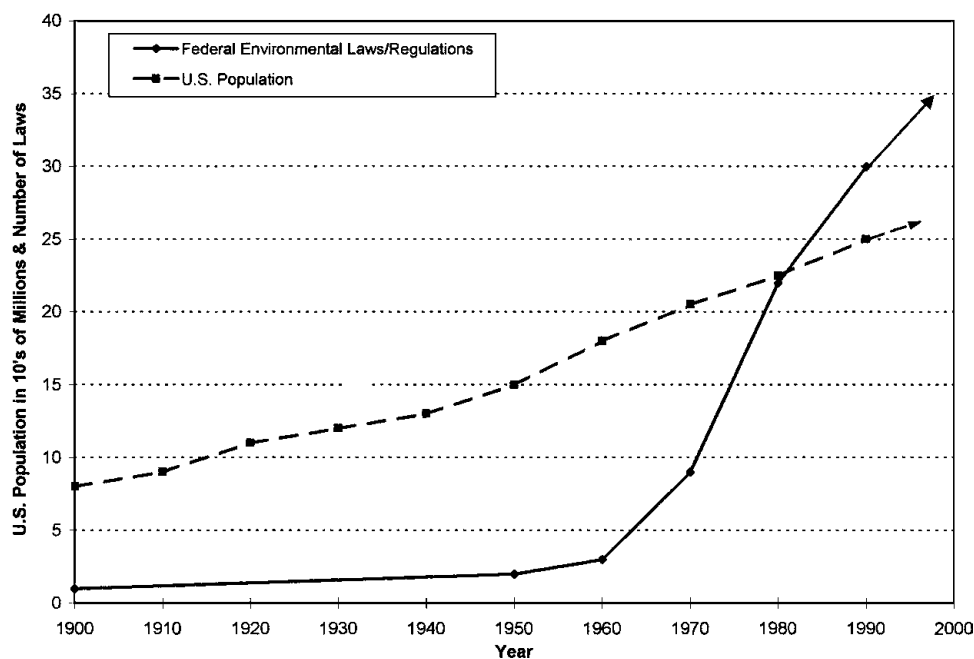


FIGURE 6 Comparison between the increase in federal environmental laws and the U.S. population with time.

times that of the United States population. In 1960, there were only 3 federal environmental laws on the books; now there are more than 30. This does not even include the much larger number of state environmental laws. Figure 6 shows both the population growth in the United States and the number of federal environmental laws and regulations as a function of time. The reality is that laws and regulations use command and control to force industry to comply.

Toward the end of the 1980s, many more industries were beginning to turn to pollution prevention as a means of avoiding the installation of expensive end-of-the pipe treatment systems. It was becoming clear to many that the succession of increasingly stringent regulations with time would ultimately lead to a complex, expensive series of treatment devices at the end of a manufacturing process, each with its own set of maintenance and performance issues.

Those industries and businesses which began to accept and implement pollution prevention solutions instead of treatment found that they not only reduced waste generation, but they also made money. As a result of these experiences, various governmental agencies began to incorporate pollution prevention requirements into new environmental laws. Congress recognized that “source reduction is fundamentally different and more desirable than waste management and pollution control,” and passed the Pollution Prevention Act in 1990.

Corporate experience has shown that the six major drivers for pollution prevention are:

1. The increasing number and scope of environmental regulations and laws.
2. Ability to save money and reduce emissions or conserve energy.
3. The rising cost and changing nature of regulations of waste treatment.
4. Greater government oversight and control of business operations.
5. More awareness by corporations in the value of pollution prevention to the business bottom line and to the customer.
6. The heightened awareness in society of the need for sustainability of the planet.

The first and second major drivers for pollution prevention, as described above, are regulations and laws and the cost of waste treatment. Extrapolation of the two curves in Fig. 6 would imply that future laws and regulations will be even more stringent and, if solved by end-of-pipe treatment, even more costly.

Figure 7 shows conceptually the cost incurred by the business to generate waste versus the amount of waste produced by a manufacturing process. Along the right-hand portion of the cost/waste curve, some processes are far to the right, whereas others are closer to the conceptual minimum. The goal of pollution prevention is to move expeditiously toward the conceptual minimum while continuing to be cost-effective. The “economic zero,” as indicated by the vertical dashed line, is the point where the slope of the curve reverses itself and normally becomes very

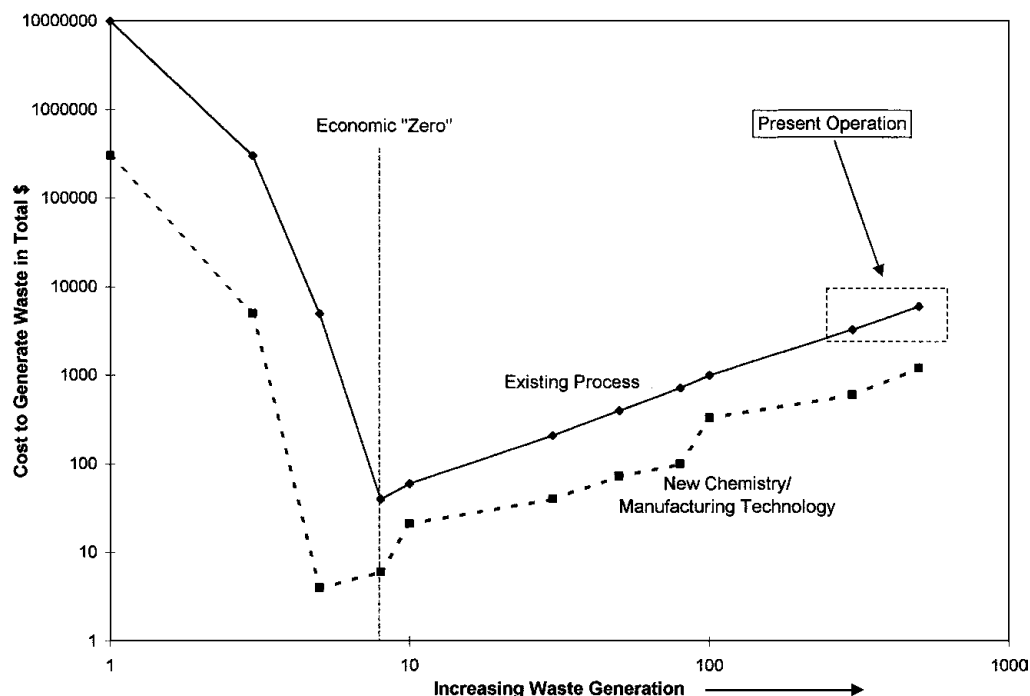


FIGURE 7 Waste generation versus business cost.

steep. Further reducing waste generation, then, requires a significantly greater capital investment, e.g., replacing large piece(s) of equipment or unit operations. Instead, to further reduce the level of waste being generated while simultaneously reducing the cost to generate this waste, new chemistry or new engineering technology is required (i.e., a new process). This is indicated by the broken curve on Fig. 7.

Federal, state, and local governments are demanding more and more information from manufacturers: not only the size, composition, and properties of waste streams that are generated, but also what chemicals are added to the process to manufacture the final product, and descriptive information on how these chemicals are used within the process. The third major driver for pollution prevention, then, becomes control of the business. When a business does not make any waste or is below a *de minimus* level, then only a minimum amount of information is required by the governing bodies; hence, business information is conserved. Thermodynamic principles govern that zero waste is not possible, and the technical challenge is develop manufacturing processes that produce minimum waste.

Figure 8 depicts schematically the degree of control business leadership has over a business versus governmental control as a function of the amount of waste being generated by a process. Normally, there will be a *de minimus* level of waste generation below which the regulations require only minimal governmental oversight, that

is, the business controls its own destiny. However, as the level of waste generation increases, so does the amount of governmental oversight. As a result, business leadership has less control of their business and is less able to respond to various business factors that might improve their bottom line. The *de minimus* point for a regulatory “zero” is normally below that for the economic “zero,” yet a business still might voluntarily choose to spend additional capital investment to increase control.

Recognizing the value of pollution prevention to the business and the customer, progressive companies are developing corporate goals to motivate their employees to reduce the amount of waste being produced. Examples include the 3M Corporate Environmental Conservation Policy and the DuPont Company’s Safety, Health and the Environment Commitment of zero waste generation and emissions, which is shown in Fig. 9.

The environmental group Grassroots Recycling Network is developing a Zero Waste Policy Paper for consumer products. The net result is that society is beginning to expect that the products and processes of the future will not generate waste and are recyclable or biodegradable.

For the businesses that have implemented pollution prevention programs, the amount saved or earned has been quite dramatic. For example, in the 3M Company, the Pollution Prevention Pays (3P) Program netted \$350 million for their U.S. plants from 1976 through 1987 while reducing waste generation by more than 425,000 tons per year.

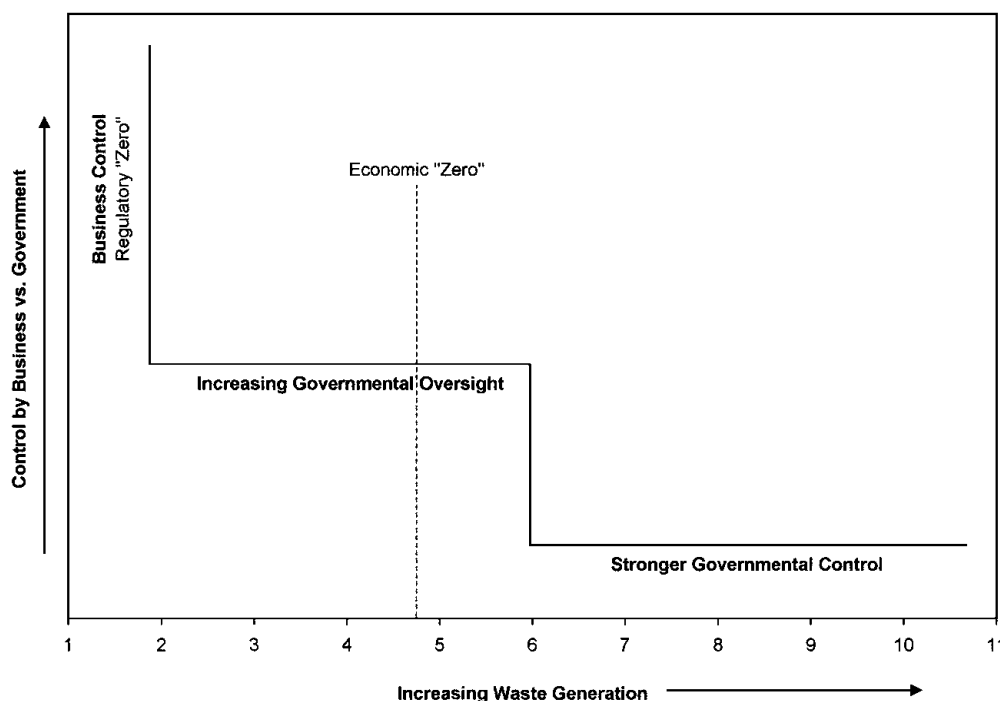


FIGURE 8 Waste generation versus business control.

A second example is the joint EPA/DuPont Chambers Works Waste Minimization Project, which resulted in a savings of \$15 million per year for only \$6.3 million in capital investment and led to a 52% reduction in waste generation.

The DuPont Company has also instituted a corporate Environmental Excellence Award program. Of the typical 550 submissions per year, approximately 70 pass the first screening and 12 are finally selected as winners. For the years 1994–1996, more than \$200 million per year positive return and \$320 million in avoided capital expenditures was realized for the 210 programs that passed the first screening.

The fifth main driver for pollution prevention, which is growing in importance, is sustainability, i.e., building a sustainable global economy or an economy that the planet is capable of supporting indefinitely. Pollution prevention is one of three ways that a company can move toward sustainability. A second way is product stewardship, where a manufactured product has minimal impact on the environment during the full manufacturing life cycle. A third step toward sustainability is through clean technology, that is, technology which has a minimum impact on the environment. Examples include (1) avoiding the use and manufacture of toxic, persistent, or bioaccumulative compounds and (2) replacing high-temperature and high-pressure processes with biotechnology routes which can manufacture products at ambient conditions.

VI. THE RECIPE FOR SUCCESS

After participating in over 75 waste reduction or treatment programs, one thing has become clear—there is a recipe for success. We have found that successful pollution prevention programs are characterized by the following four success factors:

1. Commitment by business leadership to support change and provide resources.
2. Early involvement of all stakeholders in the process.
3. Quick definition of the cost for end-of-pipe treatment, which subsequently becomes the incentive for more cost-effective pollution prevention solutions.
4. Definition and implementation of pollution prevention engineering practices and technologies that improve the business' bottom line.

The “path to pollution prevention” chart shown in Fig. 10 brings together the essential ingredients for a successful pollution prevention program, whether large or small. The core pollution prevention program or methodology is shown in the center column, and consists of three phases: the chartering phase, assessment phase, and implementation phase. The other boxes in Fig. 10 (shown with dotted lines) outline supporting information, tools, and activities that are essential to the success of the program. In many ways, these help to expedite the completion

The DuPont Commitment

Safety, Health and the Environment

We affirm to all our stakeholders, including our employees, customers, shareholders and the public, that we will conduct our business with respect and care for the environment. We will implement those strategies that build successful businesses and achieve the greatest benefit for all our stakeholders without compromising the ability of future generations to meet their needs.

We will continuously improve our practices in light of advances in technology and new understandings in safety, health and environmental science. We will make consistent, measurable progress in implementing this Commitment throughout our worldwide operations. DuPont supports the chemical industry's Responsible Care® and the oil industry's Strategies for Today's Environmental Partnership as key programs to achieve this Commitment.

Highest Standards of Performance, Business Excellence

We will adhere to the highest standards for the safe operation of facilities and the protection of our environment, our employees, our customers and the people of the communities in which we do business.

We will strengthen our businesses by making safety, health and environmental issues an integral part of all business activities and by continuously striving to align our businesses with public expectations.

Goal of Zero Injuries, Illnesses and Incidents

We believe that all injuries and occupational illnesses, as well as safety and environmental incidents, are preventable, and our goal for all of them is zero. We will promote off-the-job safety for our employees.

We will assess the environmental impact of each facility we propose to construct and will design, build, operate and maintain all our facilities and transportation equipment so they are safe and acceptable to local communities and protect the environment.

We will be prepared for emergencies and will provide leadership to assist our local communities to improve their emergency preparedness.

Goal of Zero Waste and Emissions

We will drive toward zero waste generation at the source. Materials will be reused and recycled to minimize the need for treatment or disposal and to conserve resources. Where waste is generated, it will be handled and disposed of safely and responsibly.

We will drive toward zero emissions, giving priority to those that may present the greatest potential risk to health or the environment.

Where past practices have created conditions that require correction, we will responsibly correct them.

Conservation of Energy and Natural Resources, Habitat Enhancement

We will excel in the efficient use of coal, oil, natural gas, water, minerals and other natural resources.

We will manage our land to enhance habitats for wildlife.

Continuously Improving Processes, Practices and Products

We will extract, make, use, handle, package, transport and dispose of our materials safely and in an environmentally responsible manner.

We will continuously analyze and improve our practices, processes and products to reduce their risk and impact through the product life cycle. We will develop new products and processes that have increasing margins of safety for both human health and the environment.

We will work with our suppliers, carriers, distributors and customers to achieve similar product stewardship, and we will provide information and assistance to support their efforts to do so.

Open and Public Discussion, Influence on Public Policy

We will promote open discussion with our stakeholders about the materials we make, use and transport and the impacts of our activities on their safety, health and environments.

We will build alliances with governments, policy makers, businesses and advocacy groups to develop sound policies, laws, regulations and practices that improve safety, health and the environment.

Management and Employee Commitment, Accountability

The Board of Directors, including the Chief Executive Officer, will be informed about pertinent safety, health and environmental issues and will ensure that policies are in place and actions taken to achieve this Commitment.

Compliance with this Commitment and applicable laws is the responsibility of every employee and contractor acting on our behalf and a condition of their employment or contract. Management in each business is responsible to educate, train and motivate employees to understand and comply with this Commitment and applicable laws.

We will deploy our resources, including research, development and capital, to meet this Commitment and will do so in a manner that strengthens our businesses.

We will measure and regularly report to the public our global progress in meeting this Commitment.

FIGURE 9 The DuPont commitment to safety, health, and the environment.

of the program and increase the likelihood of choosing the best options to improve the process and reduce waste generation.

The dotted boxes on the right-hand side of Fig. 10 show the information and tools available to help jump start, maintain, and increase the effectiveness of the pollution prevention program. These include:

1. How to quickly estimate the incentive for pollution prevention.
2. Generalized pollution prevention technologies and practices that apply across different industries.
3. A shortcut economic evaluation method to quickly screen the better options.

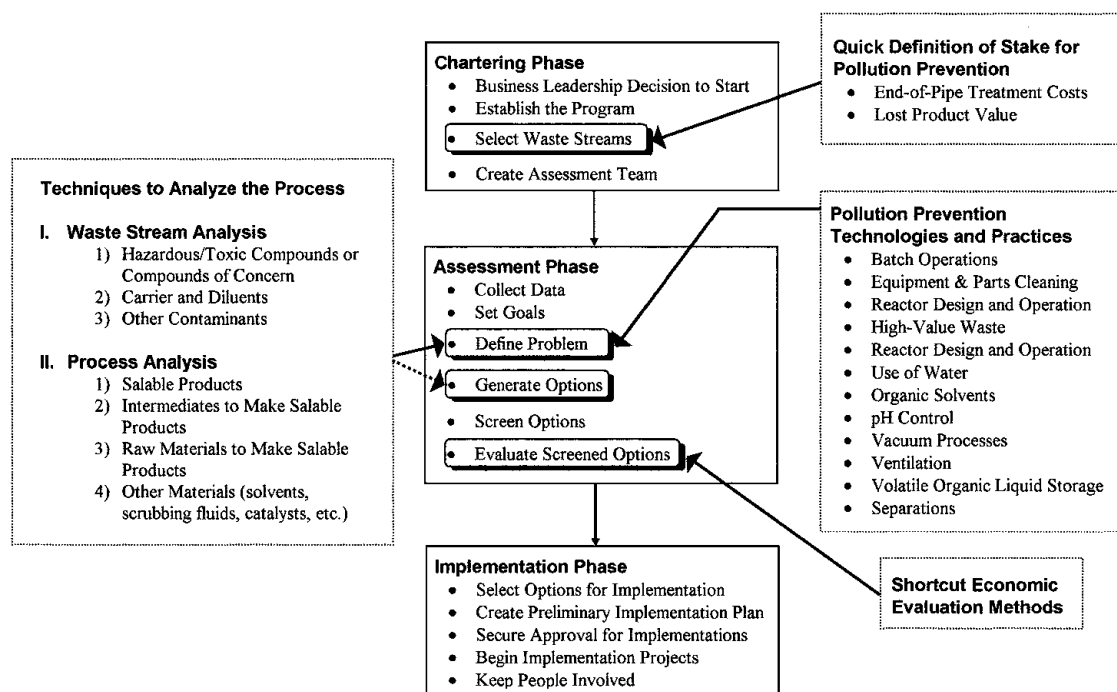


FIGURE 10 The path to pollution prevention.

The left-hand side of Fig. 10 describes two techniques to divide the waste generation problem into smaller, comprehensible parts: a waste stream analysis and a process analysis. These two analysis techniques are used to help better define the problem as well as to focus energy on the true source of the waste generation problem. The first technique, waste stream analysis, is based on the premise that most waste streams contain a carrier, such as water or air, that drives end-of-pipe treatment costs, and compound(s) or contaminants of concern that drive the need to treat the stream.

Meanwhile, the second technique, process analysis, is based on the assumption that most processes contain (1) valuable compounds and molecules that result in a saleable product (i.e., products, intermediates to make the products, and raw materials to make the intermediates/products) and (2) other compounds that add to the cost of manufacturing, which includes waste treatment costs.

VII. PROGRAM ELEMENTS

The path to pollution prevention shown in Fig. 10 is applicable at all phases of a project. In most cases, the methodology has been applied at the plant level. However, the same methodology can be used when a process is first conceived in the laboratory and at periodic intervals through startup and normal plant operation.

A. Chartering Phase

This initial phase of the pollution prevention program consists of four steps: securing business leadership support, establishing the program, selecting the waste streams, and creating a core assessment team.

1. Business Leadership Decision to Start

The decision to begin a pollution prevention program can be triggered by one or more of the drivers listed below:

- Legal requirement, i.e., state or federal regulations.
- Public image and societal expectations. This may be fueled by an adversarial attitude in the community toward the facility or process or the desire to lead the environmental movement instead of being pushed.
- Large incentive for reducing new capital investment in end-of-pipe treatment.
- Significant return by reducing manufacturing costs.
- Need to increase revenues from existing equipment.
- Corporate goal.

2. Establishing the Program

This task helps prepare the plant or manufacturing area for a successful pollution prevention effort. A key aspect of this task is to have a team leader for the program.

3. Selecting the Waste Streams

A typical process generates several major waste streams and many minor ones. The goal should be to select one or more of the major streams for the first round of waste assessments. If successful with these major streams, additional waste streams can be targeted, including minor ones, in a second round of assessments.

4. Creating an Assessment Team

In this step, a core team is selected which consists of four to six people who are best able to lead the program, perform the waste assessments, and implement the recommended process improvements. At smaller and medium-sized facilities a single individual may undertake the bulk of the pollution prevention tasks, or consultants can be used.

B. Assessment Phase

The assessment phase in many ways represents the heart of the pollution prevention program. It also tends to be where many engineers and scientists find the most enjoyment and personal satisfaction. For this reason, there is always a tendency to bypass the “softer” chartering phase and jump right into the assessment phase. This is generally a mistake. We consistently find that programs that bypass the chartering phase fail. This is because they fail to incorporate the first two major success factors listed in the recipe for success: obtaining commitment from business leadership to support change and provide resources and seeking the early involvement of all stakeholders in the process. These two major success factors arise from the chartering phase itself. However, it is also recognized that in some cases the successful implementation of the assessment phase on small projects by one or two champions could earn subsequent commitment by the business leadership for larger projects. Each company has a characteristic style for undertaking change, and the champions need to utilize these methods to accomplish their pollution prevention goals.

The assessment phase consists of tasks which help the team to understand how the target waste streams are generated and how these wastes can be reduced at the source or eliminated.

1. Collect Data

The amount of information to collect will depend on the complexity of the waste stream and the process that generates it. Material balances and process flow diagrams are a minimum requirement for most pollution prevention assessments.

2. Set Goals

This task helps the team to analyze the drivers for pollution prevention and to develop the criteria necessary to screen the options generated during the brainstorming session.

3. Define Problem

The team begins to understand the targeted waste streams and the processes that generate these streams. The waste stream and process analyses techniques are used in this step to facilitate understanding of the problem.

4. Generate Options

When the team has developed a good understanding of the manufacturing process and the source and cause of each waste stream, it should convene to brainstorm for ideas to reduce the generation of these materials.

5. Screen Options

In a separate meeting, the core assessment team will revisit the options generated during the brainstorming process to reduce the number of credible ideas carried forward.

6. Evaluate the Screened Options

More detailed engineering and economic evaluations are performed on the screened options to select the best option(s) to implement.

C. Implementation Phase

The goal of this phase is to turn the preferred options identified by the team into actual projects that reduce waste generation and emissions. Options are first selected for implementation. This should be a natural follow-up to the screening and evaluation stages described above. Next, the team needs to develop an implementation plan that includes resource requirements (both people and money) and a project timeline. This is one of the reasons that having a project engineer on the core assessment team is valuable. Third, the team must secure approval and begin project implementation. Often, this step will be according to customary local practice. Finally, people need to be kept involved throughout the entire pollution prevention program. The team leader should always be working to build and maintain momentum.

VIII. THE INCENTIVE FOR POLLUTION PREVENTION

There are several ways to determine the incentive for pollution prevention. The choice will depend on particular circumstances; that is, does a waste treatment or abatement system already exist or is a new treatment or abatement system required? Three approaches to determine the incentive for pollution prevention are described below. They are the incentive based on new end-of-pipe treatment, raw material costs, and cost of manufacture. Each of these approaches is discussed in detail below.

A. New End-of-Pipe Treatment

Gaseous and aqueous waste streams often require capital investment for new facilities or an upgrade of existing equipment, e.g., replacing an in-ground wastewater treatment basin with an aboveground treatment system in tanks. Solid wastes (both hazardous and nonhazardous) are normally handled with existing investment (e.g., site hazardous waste incinerator) or shipped off-site for disposal. In the latter case, commercial disposal costs (including the cost of transportation) serve as the incentive for pollution prevention.

1. Gas Streams

A major opportunity for savings is to reduce the flow of diluent or carrier gas (often air or nitrogen) at the source. For a gas stream containing both particulates and halogenated volatile organic compounds (VOCs), the minimum capital investment to abate this stream is about \$75 per standard cubic foot per minute (scfm) of waste gas flow.

2. Wastewater Streams

Simply speaking, wastewater streams fall into one of two general categories, those that are biologically treatable and those requiring pretreatment or stand-alone nonbiological treatment (such as chemical oxidation, stripping, and adsorption). When treating dilute aqueous organic waste streams at the end of the pipe, consideration must be given to source reduction of *both* water flow and organic loading. Substantial reductions in capital investment can result by reducing water flow and contaminant loading at the source. The magnitude of these reductions will vary with technology type, hydraulic flow, and concentration; however, the *minimum incremental* capital investments for new treatment facilities are as follows:

<i>Biodegradable aqueous waste</i>	
Incentive based on hydraulic flow	\$3000 per each additional gallon per minute (gpm)
Incentive based on organic loading	\$6000 per each additional pound organic per hour (lb/hr)
<i>Nonbiodegradable aqueous waste</i>	
Incentive based on hydraulic flow	\$1000 per each additional gpm
Incentive based on organic loading	Some technologies are sensitive to organic loading and some are not

B. Raw Materials Cost

Waste stream composition and flow rate can be used to estimate the amount of raw materials lost as waste. The product of the amount lost to waste and the purchase price sets the incentive for pollution prevention in terms of raw material cost alone.

C. Cost of Manufacture

The cost of manufacture includes all fixed and variable operating costs for the facility, including the cost for raw materials. The cost of manufacture should be cast in the form of dollars per pound (\$/lb) of a key raw material. Another number that is readily available is the product selling price in dollars per pound of product. Depending on the state of the business—excess capacity or sold out—one of these two numbers can be used to determine the incentive for pollution prevention.

- For a business operating with excess capacity, the product of the cost of manufacture (\$/lb raw material) and the amount of raw material that goes to waste (either directly or as a by-product of reaction) sets the incentive for pollution prevention.
- For a sold-out business, every additional pound of product can be sold; therefore, the product of the product selling price and the additional amount of product that can be sold determines the incentive for pollution prevention.

IX. RESOURCES

In many respects, the best set of resources for generating waste reduction ideas consists of a business' own people. However, a business will sometimes need to bring other expertise to the table to supplement its own resources. Some examples of other resources include a brainstorming facilitator, technical specialists, outsiders or wildcards, and sources of pollution prevention ideas found in the literature.

If a person cannot be found in the business who can facilitate a brainstorming session, then a consultant, possibly someone at the local university or college, will be needed.

The outsiders or wildcards should be good chemical engineering and process chemistry generalists, and not directly associated with the process. The technology specialists should be skilled in the engineering unit operations or technology areas that are most critical to waste generation in the manufacturing process, for example, drying, particle technology, reaction engineering, pumps. Most mid-size to large companies can identify the outsiders, wildcards, and technology specialists internally. For smaller firms, sources of wildcards and technology specialists include academia, engineering consultants, and research institutes.

A wealth of information is available on pollution prevention successes across many industries; however, it is primarily packaged in the form of process-specific case histories. As a result, the information is not organized in a sufficiently generalized way so as to allow the rapid transfer of knowledge from one type of industry to another. To help the practitioners of pollution prevention—engineers and scientists—more quickly to generate ideas, this process- or industry-specific information has been transformed into generalized knowledge that can be more easily implemented by project teams and existing manufacturing facilities. The information is organized in a “unit operations” format to facilitate widespread use across different processes and industries (right-hand column of Fig. 10).

Other sources for ideas are available, many on the Internet. Some examples include the following:

- The Chemical Manufacturers Association publication, “Designing Pollution Prevention into the Process: Research, Development & Engineering,” Appendices A and B
- The “Industrial Pollution Prevention Handbook” by Harry M. Freeman
- The U.S. EPA’s Pollution Prevention Directory (published annually)
- The U.S. EPA’s Pollution Prevention Information Clearinghouse (PPIC)
- The U.S. EPA’s Office of Pollution Prevention and Toxics (OPPT)
- The U.S. EPA’s Pesticide Environmental Stewardship Program
- The U.S. EPA’s EnviroSense (envirosense) database
- Case histories in journals such as *Chemical Engineering Progress*, *Journal of Chemical Technology and Biotechnology*, *Chemical Engineering*,

Environmental Progress, *Pollution Prevention Review*, and so on

- State pollution prevention offices or centers. Many states offer services to small- and medium-sized businesses (over 13,000 case studies are available on the Internet at www.P2PAYS.org)
- Private consultants or consulting firms
- Private consortia and organizations, for example, AIChE’s Center for Waste Reduction Technology (CWRT), the Center for Clean Industrial Treatment Technology (CenCITT), and the National Center for Manufacturing Sciences (NCMS)
- Pollution prevention or waste minimization centers at universities, for example, the UCLA Center for Clean Technology, the Pollution Prevention Research Center at North Carolina State University, and the Emission Reduction Research Center at the New Jersey Institute of Technology (NJIT),
- Numerous other Internet sites, such as those of the Great Lakes Pollution Prevention Centre in Canada and the Pacific Northwest Pollution Prevention Resource Center.

A review on using the Internet for pollution prevention was published by [Scott Butner \(1997\)](#) at the Battelle Seattle Research Center. All of these resources can be used to help prepare a brainstorming team for the generation of ideas.

X. ENGINEERING EVALUATIONS OF THE PREFERRED OPTIONS

Engineering evaluation is the application of a full range of engineering skills to business decision making. It aids decision making by translating technical options into economic impact, guidance that is fundamental to business decisions. The evaluation quickly focuses on only those data and analyses which are essential to quantify technical and economic feasibility. For each preferred option, the evaluation involves the following:

- Defining the commercial process
- Flowsheeting
- Analyzing the process
- Defining manufacturing facilities
- Estimating investment and manufacturing cost
- Analyzing economics
- Assessing risk

The evaluation provides an objective view for decision making that is grounded in both engineering science and economics.

XI. WASTE STREAM AND PROCESS ANALYSES

Properly defining and subdividing the problem ultimately leads to the best pollution prevention solutions. The goal is to frame the problem such that the pertinent questions arise. When the right questions are asked, the more feasible and practical solutions for pollution prevention become obvious. Analyzing the manufacturing process in this manner before and during the brainstorming session will often result in an improved process that approaches an absolute minimum in waste generation and emissions.

A. Waste Stream Analysis

The best pollution prevention options cannot be implemented unless these are identified. To uncover the best options, each waste stream analysis should follow four steps:

1. List all components in the waste stream, along with any key parameters. For instance, for a wastewater stream these could be water, organic compounds, inorganic compounds (both dissolved and suspended), pH, etc.
2. Identify the compounds triggering the concern, for example, compounds regulated under the Resource Conservation and Recovery Act (RCRA), hazardous air pollutants (HAPs), and carcinogenic compounds. Determine the sources of these compounds within the process. Then develop pollution prevention options to minimize or eliminate the generation of these compounds.
3. Identify the highest volume materials (often these are diluents, such as water, air, a carrier gas, or a solvent) because these materials or diluents often control the investment and operating costs associated with end-of-pipe treatment of the waste streams. Determine the sources of these diluents within the process. Then develop pollution prevention options to reduce the volume.
4. If the compounds identified in step 2 are successfully minimized or eliminated, identify the next set of compounds that has a large impact on investment and operating cost (or both) in end-of-pipe treatment. For example, if the aqueous waste stream was originally a hazardous waste and was incinerated, eliminating the hazardous compound(s) may permit the stream to be sent to the wastewater treatment facility. However, this may overload the biochemical oxygen demand (BOD) capacity of the existing wastewater treatment facility. If so, it may be necessary to identify options to reduce organic load in the aqueous waste stream.

B. Process Analysis

In the manufacturing facility in Fig. 2 all of the materials added to or removed from the process are valuable to the business. Therefore, to help frame the problem for a real manufacturing facility, a process analysis should be completed.

For either a new or existing process, the following steps are taken:

1. List all raw materials reacting to saleable products, any intermediates, and all salable products. This is "list 1."
2. List all other materials in the process, such as nonsaleable by-products, solvents, water, air, nitrogen, acids, bases, and so on. This is "list 2."
3. For each compound in list 2, ask "How can I use a material from List 1 to do the same function of the compound in list 2?" or "How can I modify the process to eliminate the need for the material in list 2?"
4. For those materials in list 2 that are the result of producing nonsaleable products (i.e., waste by-products), ask "How can the chemistry or process be modified to minimize or eliminate the wastes (for example, 100% reaction selectivity to a desired product)?"

Analyzing the process in these ways and then applying fundamental engineering and chemistry practices will often result in a technology plan for driving toward a minimum waste generation process. Other key ingredients for a successful pollution prevention program are a proven methodology and the ingenuity of a savvy group of people to generate the options.

XII. WHEN SHOULD ONE DO POLLUTION PREVENTION?

The continuum depicted in Fig. 11 shows the relative merits of when a pollution prevention program should be implemented. The decision of how far to move toward the lowest waste and emissions design will depend on a number of factors including corporate and business environmental goals, regulatory pressures, economics, the maturity of the process, and product life. It is safe to say, "the earlier, the better." If one can make changes during the R&D stage of the process or product life cycle, then one has the best opportunity to make significant reductions in waste generation at the source. However, as one moves down the continuum from R&D through process design and engineering and post-startup operation, one's

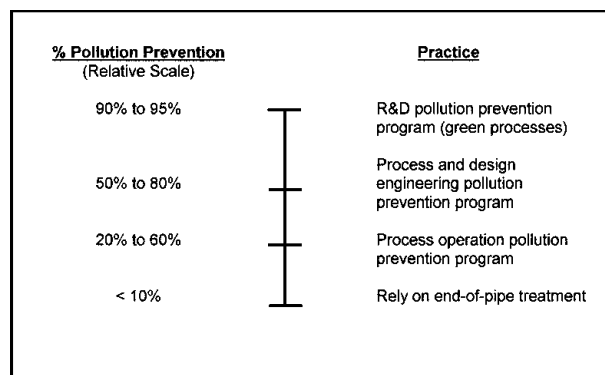


FIGURE 11 Pollution prevention methodology continuum.

dependence on end-of-pipe treatment grows. At the bottom of the continuum is a total reliance on end-of-pipe treatment. Here, pollution prevention may be manifested in the form of energy savings or a reduction in air flow to the abatement device, etc.

A. Pollution Prevention during Research and Development

Research and development programs typically progress through three distinct phases: process conception, laboratory studies, and pilot plant testing. The level of effort and detail required in pollution prevention assessment depends on the particular R&D phase. Generally speaking, studies are qualitative during process conception, semiquantitative in laboratory studies, and quantitative in pilot plant testing. The basic steps in a pollution prevention study, however, are the same in each phase.

During process conception, reaction pathways, inherent process safety, general environmental impacts of products, and waste streams are studied, and pollution prevention concepts are formulated.

During laboratory studies, reaction chemistry is confirmed, waste streams are characterized, process variables are tested, pollution prevention options are identified, data are collected for the pilot plant and process design, and the potential impact of environmental regulations is determined.

During pilot plant studies, laboratory results are confirmed, process chemistry is finalized, key process variables are tested, equipment design is evaluated, and waste characteristics are defined. It is especially important at this stage of R&D that all major environmental cost areas are understood as these relate to the overall viability of a commercial project.

B. Pollution Prevention during Process and Design Engineering

While the greatest opportunity for cost-effective waste reduction at the source exists at the R&D stage, additional

opportunities may exist during process engineering and should be explored. The potential to reduce waste and pollutant releases in this stage is impacted by the selection of process configuration (batch versus continuous, for example), process conditions (such as temperature and pressure), manufacturing procedures, design and selection of processing equipment, and process control schemes.

As a project moves into the detailed design stage (sometimes referred to as the “mechanical design stage” or “production design”), source reduction opportunities typically diminish. The main reason is that the process and preliminary plant design become fixed and the project becomes schedule-driven. The focus at this stage shifts from the chemical process to equipment and facility design. The emphasis at this point should be to protect groundwater from spills and to minimize or eliminate fugitive emissions.

C. Pollution Prevention during Process Operation

If the pollution prevention program began during the research stage, then a pollution prevention analysis is not necessary until 3 years after startup of the process. Ideally, a pollution prevention program should be completed every 3–5 years.

For a process that does not have a history of doing pollution prevention, a pollution prevention program can generally realize a greater than 30% reduction in waste generation and a greater than 20% reduction in energy usage.

XIII. CASE STUDIES

Four case studies are presented below which exemplify the role of the structured pollution prevention program methodology, the value of quickly defining the incentive for pollution prevention using the cost of end-of-pipe treatment, and the benefits of using the waste stream and process analyses to parse the problem at hand. Five case studies are also presented illustrating pollution prevention results at each of the stages described in Fig. 11.

A. Program Elements—U.S. EPA and DuPont Chambers Works Waste Minimization Project

In May 1993, the U.S. EPA and DuPont completed a joint 2-year project to identify waste reduction options at the DuPont Chambers Works site in Deepwater, New Jersey. The project had three primary goals as conceived:

1. Identify methods for the actual reduction or prevention of pollution for specific chemical processes at the Chambers Works site.

2. Generate useful technical information about methodologies and technologies for reducing pollution, which could help the U.S. EPA assist other companies implementing pollution prevention/waste minimization programs.
3. Evaluate and identify potentially useful refinements to the U.S. EPA and DuPont methodologies for analyzing and reducing pollution and/or waste generating activities.

The business leadership was initially reluctant to undertake the program, and was skeptical of the return to be gained when compared against the resources required. After completing a few of the projects, however, the business leadership realized that the methodology identified revenue-producing improvements with a minimum use of people resources and time, both of which were in short supply.

The pollution prevention program assessed 15 manufacturing processes and attained the following results:

- A 52% reduction in waste generation.
- Total capital investment of \$6,335,000.
- Savings and earnings amounting to \$14,900,000 per year.

Clearly, this is a very attractive return on investment, while also cutting waste generation in half. No matter which methodology was used, the EPA's or DuPont's, the results were the same. The key to the site's success was following a structured methodology throughout the project and allowing creative talents to shine in a disciplined way.

B. Incentive for Pollution Prevention—Gas Flow Rate Reduction

A printing facility in Richmond, Virginia, uses rotogravure printing presses to produce consumer products packaging materials. Typical solvents used are toluene, isopropyl acetate, acetone, and methyl ethyl ketone. Driven by the U.S. EPA's new source performance standards for the surface coating industry, the site installed a permanent total enclosure (PTE) around a new press so as to attain a 100% VOC capture efficiency. Leaks from the hot air convection dryers and other fugitive emissions from the coating operation are captured in the press enclosure and routed, along with the dryer exhaust, to a carbon adsorber for recovery. Overall VOC removal efficiency for the enclosure and recovery system is greater than 95%. While many rotogravure press installations use the total pressroom as the enclosure, this facility was one of the first to install a separate, smaller enclosure around the new press. Notable features of the enclosure include the following:

- Quick-opening access doors
- A dryer which serves as part of the enclosure to minimize the enclosure size
- VOC concentration monitors which control air flow to each dryer stage to maintain the dryers at 25–40% of the LEL (lower explosive limit)
- Damper controls which maintain a constant exhaust rate from the enclosure to ensure a slight vacuum within the enclosure.

If the pressroom had been used as the enclosure, the amount of ventilation air requiring treatment would have been close to 200,000 scfm. Instead, the use of the enclosure and the LEL monitors reduced the air flow to the adsorber to 48,000 scfm. This resulted in an investment savings for the carbon adsorber of approximately \$5,000,000. The installed cost of the 1700-ft² enclosure was only \$80,000, or \$47/ft². Knowing the investment required to treat the entire 200,000 scfm provided a clear incentive for the business to reduce air flow at the source through segregation.

C. Waste Stream Analysis—Nonaqueous Cleaning

In a sold-out market, a DuPont intermediates process was operating at 56% of peak capacity. The major cause of the rate limitation was identified as poor decanter operation. The decanter recovered a valuable catalyst, and the poor operation was caused by fouling from catalyst solids. Returning the process to high utility required a 20-day shutdown. During the shutdown, the vessel was pumped out and cleaned by water washing. The solids and hydrolyzed catalyst were then drummed and incinerated. A waste stream analysis identified three cost factors: the volume of wastewater that had to be treated, the cost of the lost catalyst, and the incineration cost.

An analysis of the process and ingredients indicated that the decanter could instead be bypassed and the process run at a reduced rate while the decanter was cleaned. A process ingredient was used to clean the decanter, enabling recovery of the catalyst (\$200,000 value). The use of the process ingredient in place of water cut the cleaning time in half, and that, along with continued running of the process, eliminated the need to buy the intermediate on the open market. The results were a 100% elimination of a hazardous waste (125,000 gallons per year) and an improved cash flow of \$3,800,000 per year.

D. Process Analysis—Replace Solvent with a Process Intermediate, Product, or Feed

At a DuPont site, organic solvents used in the manufacture of an intermediate monomer were incinerated as a

hazardous waste. These organic solvents were used to dissolve and add a polymerization inhibitor to the process. Alternative nonhazardous solvents were considered and rejected because these solvents would not work in the existing equipment. However, with the help of process analysis techniques, the intermediate monomer was found to have the same dissolution capacity as the original organic solvents. As a result, the site replaced the organic solvents with the intermediate monomer. By utilizing existing equipment, realizing savings in solvent recovery, and reducing operating and incineration costs, the project achieved a 33% internal rate of return (IRR) and a 100% reduction in the use of the original solvents.

E. R&D Phase

1. Waste Reduction through Control of the Reaction Pathway

In hydrocarbon oxidation processes to produce alcohol, there is always a degree of overoxidation. The alcohol is often further oxidized to waste carboxylic acids and carbon oxides. If boric acid is introduced to the reactor, the alcohol reacts to form a borate ester, which protects the alcohol from further oxidation. The introduction of boric acid terminates the by-product formation pathway and greatly increases the product yield. The borate ester of alcohol is then hydrolyzed, releasing boric acid for recycle back to the process. This kind of reaction pathway control has been applied to a commercial process, resulting in about a 50% reduction in waste generation once the process was optimized.

2. Waste Reduction through Catalyst Selection

For chemical processes involving catalysis, proper selection of catalysts can have a major impact on product formation. One example is the ammoxidation of propylene to form acrylonitrile. Different catalysts result in a wide range of product and by-product yields. By-product yields of 50–80% (based on carbon) have been reported in the literature. Use of a different catalyst provided a 50% reduction in waste generation by increasing product yield from 60% to 80%.

F. Process and Design Engineering Phase

1. Reuse Reaction Water in Wash Step

A dehydration reaction generates a continuous stream of water, which requires disposal. A separate product wash step uses deionized water, which is also disposed. Testing verified that the dehydration water could replace the deionized water in the wash step without product qual-

ity impacts. Initial concerns about product quality were unfounded. Total waste generation was reduced by the quantity of dehydration water which is reused.

2. Groundwater Protection

At a grassroots facility, one company utilized a groundwater protection strategy which included several construction tactics not required by current environmental regulations. Chemical storage tanks were designed with double bottoms to allow leak detection before environmental damage. Similarly, one nonhazardous process water pond was constructed with synthetic liners to eliminate the possibility of groundwater impact from any pollutants. Nonhazardous process water ditches, traditionally used in chemical plants, were replaced with hard-piped sewer lines to eliminate the leak potential inherent with concrete.

G. Existing Process Operation

1. Reduced Hazardous Waste Generation

At a chemical manufacturing site, a series of distillation columns are used to purify different product crudes in separate campaigns. At the conclusion of each campaign, a portion of product crude was used to wash out the equipment. When the crude became too contaminated, it was sent for destruction in a hazardous waste incinerator. First, an analysis of the washing procedure of a decant tank indicated that only 1/10 of the product crude wash material was really needed to effect cleaning. Second, a dedicated pipeline for each crude was installed, thus eliminating the need to flush the line between campaigns. Third, an extended and improved drainage procedure was developed for a large packed-bed distillation column. Finally, the product specifications were relaxed, so that fewer washes were required to maintain product specifications. Capital investment for these process changes was \$700,000; however, the project had a positive net present value of more than \$3 million, and realized a 78% reduction in waste generation.

XIV. CONCLUSION

Pollution prevention is becoming an integral part of business operations, both new and existing. As the drive toward a more sustainable human society strengthens, pollution prevention will become even more necessary for a business to survive. There are large opportunities to do both pollution prevention and improve the economic return on manufacturing processes. Everyone in a business can contribute to reducing manufacturing waste. In this article we described pollution sources, pollution prevention

techniques, and how everyone can contribute to pollution prevention.

SEE ALSO THE FOLLOWING ARTICLES

ENVIRONMENTAL TOXICOLOGY • HAZARDOUS WASTE INCINERATION • POLLUTION, AIR • POLLUTION CONTROL • POLLUTION, ENVIRONMENTAL • RADIOACTIVE WASTE DISPOSAL • SOIL AND GROUNDWATER POLLUTION • TRANSPORT AND FATE OF CHEMICALS IN THE ENVIRONMENT • WASTE-TO-ENERGY SYSTEMS • WASTE-WATER TREATMENT AND WATER RECLAMATION • WATER POLLUTION

BIBLIOGRAPHY

- 3M Corporation. (1983). "The 3P Program," 3M, St. Paul, MN.
- Bringer, R. P. (1989). "Pollution prevention program saves environment and money," *Adhesives Age* **32**, 33–36.
- Butner, S. (1997). "Using the Internet for pollution prevention," *Pollution Prevention Rev.* **7**(4), 67–74.
- Chemical Manufacturers Association. (1993). "Designing Pollution Prevention into the Process: Research, Development and Engineering," Chemical Manufacturers Association, Washington, DC.
- Dyer, J. A., and Mulholland, K. L. (1994). "Toxic air emissions: What is the full cost to your business?" *Chem. Eng. Environ. Eng. Spec. Suppl.* **101**(2), 4–8.
- Dyer, J. A., and Taylor, W. C. (1994). "Waste management: A balanced approach." In "Proceedings of the Air and Waste Management Association's 87th Annual Meeting and Exhibition," 94-RP122B.05, Cincinnati, OH.
- Freeman, H. M. (1995). "Industrial Pollution Prevention Handbook," McGraw-Hill, New York.
- Hart, S. L. (1997). "Beyond greening: Strategies for a sustainable world," *Harvard Bus. Rev.* **75**(1), 66–76.
- Mulholland, K. L., and Dyer, J. A. (1999). "Pollution Prevention: Methodology, Technologies and Practices," American Institute of Chemical Engineers, New York.
- North Carolina Office of Waste Reduction. <http://www.P2PAYS.org>.
- Overcash, M. (1987). "Techniques for Industrial Pollution Prevention," Lewis, Chelsea, MI.
- Overcash, M. (1991). "Assistance in Development of the EPA Program for Pollution Prevention: The Distinguished Visiting Scientist Report," Risk Reduction Engineering Research Laboratory, Cincinnati, OH.
- Overcash, M. (1992). "Pollution prevention in the United States, 1976–1991." In at "Cleaner Production and Waste Minimization, London, Oct. 22–23, 1992," IBC, London.
- Overcash, M., and Miller, D. (1981). "Integrated Hazardous Waste Management Today Series," American Institute of Chemical Engineers, New York.
- Thurber, J., and Sherman, P. (1995). Pollution prevention requirements in United States environmental laws. In "Industrial Pollution Prevention Handbook" (H. M. Freeman, ed.), pp. 27–49, McGraw-Hill, New York.
- U.S. Environmental Protection Agency. (1992). "Facility Pollution Prevention Guide," EPA/600/R-92/088, U.S. EPA, Office of Research and Development, Washington, DC.
- U.S. Environmental Protection Agency. (1993). "DuPont Chambers Works Waste Minimization Project," EPA/600/R-93/203. U.S. EPA, Office of Research and Development, Washington, DC.



Pulp and Paper

Raymond A. Young

University of Wisconsin-Madison

Robert Kundrot

Koppers Company

David A. Tillman

*Envirosphere Company, A Division of Ebasco Services
Incorporated*

- I. Introduction
- II. Furnish for Pulp and Paper
- III. Chemical Pulping
- IV. Mechanical Pulping of Wood
- V. Bleaching of Wood Pulps
- VI. Papermaking
- VII. Recycling in Pulp and Paper

GLOSSARY

Alpha-cellulose Alpha-cellulose, also known as chemical cellulose, is a highly refined, insoluble cellulose from which all sugars, pectin, lignin, and other soluble materials have been removed. It is commonly used in the production of nitrocellulose, carboxymethylcellulose, dissolving pulps, and other compounds.

Bleaching Chemical process in pulping that removes or alters the remaining lignin after the pulping process and improve the brightness and stability of the pulp.

Boxboard General term designating the paperboard used for fabricating boxes. It may be made of wood pulp or paper stocks or any combinations of these and may be plain, lined, or clay coated. Terminology used to classify boxboard grades is normally based

upon the composition of the top liner, filler, and back liner.

Burst strength Measure of the ability of a sheet to resist rupture when pressure is applied to one of its sides by a specified instrument, under specific conditions. A burst factor is obtained by dividing the burst strength in grams per square centimeter by the basis weight of the sheet in grams per square meter.

Cellulose Cellulose is the main polysaccharide in living plants and trees, forming its skeletal structure. Cellulose is a polymer of B-D-glucose with an approximate degree of polymerization (DP) from 2000 to 4000 units.

Cord Measure of roundwood or pulpwood representing a stack of such wood 4 ft × 4 ft × 8 ft or 128 ft³.

Dissolving pulp Dissolving pulps are also referred to as chemical cellulose. This pulp is taken into solution

to make cellulosic products such as rayon, cellulose acetate, and nitrocellulose. These pulps are high alpha-cellulose pulps containing a minimum of hemicelluloses, lignin, and extractives depending on grade.

Fourdrinier screen (or wire) Endless belt woven of wire suitable for use on the fourdrinier machine on which pulp fibers are felted into paper or paperboard.

Furnish This is the mixture, and proportion thereof, of fibrous and other materials being conditioned or prepared for the paper machine. It is also to refer to the materials being put together.

Hemicellulose Group of carbohydrates found in the cell wall in more or less intimate contact with cellulose. The hemicelluloses are more soluble than cellulose and much more readily hydrolyzed into sugars.

Holocellulose Total carbohydrate fraction of wood remaining after the removal of lignin and solvent extractable substances.

Lignin One of the principal constituents of woody cell walls, whose exact chemical composition is still unknown. In general lignin is aromatic or hydroaromatic in nature containing phenyl-propane units and lacking fused polycyclic hydrocarbons such as naphthalene or anthracene. Lignin is sometimes considered to be the “glue” holding wood fibers together.

Paperboard One of the two broad subdivisions of paper (general term), the other being paper (specific term). The distinction between paperboard and paper is not sharp but broadly speaking, paperboard is heavier in basis weight, thicker, and more rigid than paper. In general, all sheets thicker than .012 in. are classified as paperboard.

Paper machine Machine on which paper or paperboard is manufactured. The most common type is the fourdrinier machine using the fourdrinier wire as a felting medium for the fibers.

Tear strength (tearing resistance) Force required to tear a specimen under standardized conditions. The tearing resistance in grams (per sheet) multiplied by 100 and divided by the basis weight in grams per square meter equals the tear factor.

Wet strength Strength of a specimen of paper after it has been wetted with water under specified conditions.

THE TERM PULP is used to describe the raw material for the production of paper and allied products such as paperboard, fiberboard, and dissolving pulp for the subsequent manufacture of rayon, cellulose acetate, and other cellulose products. More specifically, pulp is wood or other biomass material that has undergone some degree of chemical or mechanical action to free the fibers either individually or as fiber bundles from an embodying matrix. Paper,

or any other allied product mentioned above, is a term used to describe pulp after a reconsolidation into sheet or board form has occurred.

I. INTRODUCTION

Pulp and paper refers to the processes employed to convert wood fiber into paper and allied products used in such applications as communications, packaging, and construction. Pulp and paper technologies or processes capitalize upon the anatomical, physical, and chemical properties of wood and, to a much lesser extent, other sources of biomass. The application of those technologies or processes has led to the development of a highly capital intensive industry with worldwide sales on the order of \$100 billion per year.

A. Dimensions of the Pulp and Paper Industry

The U.S. pulp and paper industry produces almost 100 million tonnes (metric tons) of paper annually. The paper finds its way into a wide variety of products including newsprint, tissue, printing and writing papers, unbleached kraft paper, bleached boxboard, unbleached kraft linerboard, corrugating medium, recycled paperboard, and numerous other commodities. These paper products compete with plastics in the packaging of consumer goods from eggs to milk. They are also used in sanitary applications where disposability is highly desirable.

The production of millions of tons of paper annually requires a capital intensive industry. A modern pulp and paper facility such as the Leaf River Mill shown in [Fig. 1](#) can cost in excess of \$800 million to construct. Pulp and paper manufacturing throughout the world is a vast industry, with production levels approaching 300 million tonnes/year. The dominant pulp and paper producing countries include: Canada, Sweden, Finland, Japan, Brazil, and Russia. The pulp and paper industry is typically located near convenient, low-cost sources of wood as the raw material.

B. Historical Development of the Pulp and Paper Industry

Paper has been produced since the dawn of civilization. Raw material for early papers included old paper (recycling), rags, and cotton linters. During the last half of the 19th century and the first half of the 20th century, however, a series of inventions occurred that revolutionized the pulp and paper industry. These innovations are shown in [Table I](#), and are reviewed in detail elsewhere. These developments made wood the desirable raw material



FIGURE 1 Overview of the recently completed bleached Kraft pulp mill built by Leaf River Forest Products in Mississippi. [Photo courtesy of Leaf River Corp.]

for wood pulping, and produced a range of pulp and paper products with varying strength, printability, and other characteristics.

By 1900 a sufficient technology base was established to support the growth of the pulp and paper industry. Of particular importance was the Kraft process, and Kraft pulping has become the dominant method for liberating usable fiber from wood. The domination of Kraft pulping became particularly pronounced after 1920. It was aided by the following inventions: (1) the Tomlinson furnace, permitting simultaneous energy and chemical recovery from spent

pulping liquor; (2) the Kamyr continuous digester, converting the industry from batch to continuous processes; (3) the sawmill debarker and chipper, making residues as well as cordwood available as furnish; and (4) secondary innovations such as the diffusion washer and displacement bleaching system.

The thermomechanical pulping (TMP) invention in 1939, and the subsequent introduction of this technology from 1968–1973, and refiner mechanical pulping (RMP), permitted the application of mechanical pulping systems to residue sources of wood. Their development spurred the improvement of stone groundwood (SGW) pulping by the introduction of pressurized groundwood (PGW) systems.

This article is organized first to examine the issues associated with pulp mill raw materials. It then focuses on chemical pulping, mechanical pulping, bleaching, and papermaking. It is designed to overview the major technical concerns associated with these technologies.

II. FURNISH FOR PULP AND PAPER

The dominant raw material for pulp and paper is wood either harvested specifically for pulp production or produced

TABLE I Dominant Process Inventions in the Pulp and Paper Industry^a

Year	Pulping process invented
1844	Groundwood mechanical pulping
1851	Soda pulping
1866	Sulfite pulping
1880	Semichemical pulp
1884	Kraft (sulfate) pulping
1939	Thermomechanical pulp

^aFrom Libby, C. E. (1962). "Pulp and Paper Science and Technology," Vol. 1, Pulp. McGraw-Hill, New York.

as a byproduct of lumber or plywood manufacturing. In recent years the by-product source of wood has become increasingly important, virtually displacing all cordwood in Pacific Coast pulp mills. Today, over 40% of all wood utilized by U.S. pulp mills comes from such chips. This development resulted both from technologies to produce and to utilize chips from sawmill slabs and green clippings from the plywood mill.

Although wood is the dominant raw material for pulp and paper in the developed world, a wide range of fibers are utilized for papermaking in other parts of the world. In many countries pulp production is based entirely on agro-based fibers and over 25 countries depend on agro-based fibers for over 50% of their pulp production. The leading countries for production of pulp and paper from agro-based fibers are China and India, with China having over 73% of the world's agro-based pulping capacity. China mainly utilizes straw for papermaking while India and Mexico utilize large quantities of sugar cane bagasse (fiber waste from sugar production). India also incorporates some jute fiber and large quantities of bamboo, although the supply of bamboo is not sufficient to meet demands for paper production. There has been considerable interest in the use of kenaf as an alternate fiber source in the U.S. and a number of successful press runs of kenaf based paper (82–95%) were carried out in the pressrooms of the *Bakersfield Californian*, the *Houston Chronicle*, the *Dallas Morning News* and the *St. Petersburg Times*.

Practically any natural plant can be utilized as a source of papermaking fibers, but there is considerable variation in the quality of paper realized from alternate plant sources. Factors such as fiber length, content of non-fibrous components such as parenchyma tissue, contaminants such as silica, etc. greatly influence the quality of the final sheet. Procurement of sufficient quantities of the raw material and seasonal fluctuations in supply can also pose problems. It is also necessary to use alternate pulping equipment to handle the plant materials since the material tends to mat down in the digester making it difficult to get uniform circulation of the cooking chemicals.

A. Wood Availability

The U.S. has over 200 million hectares (490 million acres or 770,000 square miles) of commercial forest land, a resource base that routinely produces more cubic meters of timber than is harvested annually. Of the timber producing regions of the United States, only the Pacific Coast witnesses more harvest than growth. The anomaly of the Pacific Coast results from the large inventory of old growth Douglas-fir. As second growth stands become more prominent, this harvest/growth deficit will be reversed. In the south, the major pulp and paper producing

region of the United States, growth routinely exceeds harvest. This situation is aided by short rotation ages of pulpable southern species from loblolly pine to American sycamore. Loblolly pine can be grown in 15- to 30-year rotations, while American sycamore can be grown in 5- to 10-year rotations.

Pulpwood also is plentiful in such countries as Canada and the Russia; and abundant tropical forests exist in such countries as Brazil. Adequate wood supplies exist in Scandinavia as well. Silvicultural practices in the Scandinavian region, coupled with intensive utilization of harvested materials, have prevented undue scarcity in that geographic area.

B. Wood Quality

Issues of quality include anatomical, physical, and chemical properties of various types of furnish. Anatomical concerns focus upon wood fiber length, because fiber length influences a variety of paper properties from strength to printability. Physical properties of consideration include various measures of strength. Measures of strength can be inferred from fiber length and specific gravity. Chemical properties of concern include percentage composition, cellulose, the hemicelluloses, and lignin. Cellulose content largely determines yield of chemical pulping. Lignin content determines the higher heating value of spent pulping liquor. The extractives content determines the economic value of byproduct production of naval stores from Kraft pulp mills. Such mills are the dominant sources of rosin, distilled tall oil, and turpentine in the current forest products industry.

Typical properties of selected wood species are shown in Table II. Note that the clear distinctions between the softwoods and hardwoods include fiber length, hence resulting pulp strength. Softwoods are clearly superior from a strength perspective. Note, also the higher cellulose content of hardwoods—implying that such species as trembling aspen will have higher chemical pulp yields than coniferous woods. In general hardwoods have 45% cellulose, 30% hemicelluloses, and 20% lignin, while softwoods will have 42% cellulose, 27% hemicelluloses, and 28% lignin. It is useful to note that properties of wood change as trees age. For example, Bendston has shown that an 11-year-old loblolly pine has a tracheid length of 2.98 mm and a cell wall thickness of 3.88 μm . A 39-year old tree of the same species will have a tracheid length of 4.28 mm and a cell wall thickness of 8.04 μm . More mature trees will yield higher strength fibers.

Given the general properties of wood furnish as identified above, it is now important to examine specific chemical and mechanical pulping, bleaching, and papermaking technologies.

TABLE II Selected Fundamental Properties of Several Wood Species^a

Species	Fiber length (mm)	Specific gravity	Moisture content		Summative chemical composition			
			(percent, O.D.)		Cellulose (percent)	Hemicelluloses (percent)	Lignin (percent)	Extractives (percent)
			Heartwood	Sapwood				
Softwoods								
Douglas-fir	5.0	0.45–0.50	37	115	38.8	26.6	29.3	5.3
Eastern hemlock	3.5	0.38–0.40	97	119	37.7	28.4	30.5	3.4
Larch	5.0	0.48–0.52	54	110	41.4	30.4	26.4	1.8
White spruce	3.5	0.37–0.40	34	128	39.5	30.9	27.5	2.1
Southern pines	4.6	0.47–0.51 ^b	33 ^b	110 ^b	42 ^c	24	27 ^c	3.5
Hardwoods								
Trembling aspen	1.25	0.35–0.39	95	113	56.6 ^c	27.1 ^c	16.3 ^c	
Red maple	1.00	0.49–0.54	65	72	42.0	29.4	25.4	3.2 ^c
Beech	1.20	0.56–0.64	55	72	39.4	34.6	24.8	1.2
Paper birch	1.20	0.48–0.55	89	72	39.4	36.7	21.4	2.6

^aFrom Sjöström, E. (1981). "Wood Chemistry: Fundamentals and Applications." Academic Press, New York; and Wenzl, H. (1970). "The Chemical Technology of Wood," Academic Press, New York.

^bValues for loblolly pine.

^cExtractive-free basis.

III. CHEMICAL PULPING

Chemical pulping consists of treating wood chips with specific chemicals in order to break the internal lignin and lignin-carbohydrate linkages and liberate pulp fibers. Chemical pulping not only liberates individual wood fibers, but also removes most of the lignin from the pulp and "flexibilizes" the fibers. Because the pulp fibers are liberated chemically rather than mechanically, the pulp contains a higher percentage of whole long fibers. Flexibility permits more contact points between individual fibers in the ultimate product—the sheet of paper. Consequently, chemical pulps are inherently stronger than pure mechanical pulps.

Chemical pulping is used to produce not only high-strength pulps but also essentially pure cellulose pulps (cellulose or dissolving pulps). The high-strength pulps are used in paper and paperboard products as discussed later. Dissolving pulps are used to produce a range of products including cellophane, cellulose acetate, carboxymethyl-cellulose (CMC), rayon, and a range of other modified cellulose products.

A. The Range of Chemical Pulping Processes

Chemical pulping has been performed or proposed with a wide variety of reactants. Today the dominant chemicals used in pulping are sulfur based, although numerous sulfur-free processes have been proposed. The processes available currently include sulfate or Kraft pulping, acid and alkaline sulfite pulping, neutral sulfite semichemical

(NSSC) pulping, and soda pulping. Of these the Kraft process has become dominant and for the following reasons: (1) it can produce useful pulps from all wood species; (2) it readily permits chemical and energy recovery from the spent pulping liquor and was the first pulping process to do so; and (3) it regularly produces the highest-strength pulps.

Because Kraft is the dominant chemical pulping method available today, it is the focus of this section. Other chemical pulping methods are presented by comparison.

B. Principles of Chemical Pulping

Chemical pulping dissolves the lignin from the middle lamella in order to permit easy fiber liberations. Not all of the lignin is removed, however, since 3–10% by weight remains in the pulp depending upon wood species and pulp properties desired.

1. Kraft Pulping

In Kraft pulping, dissolution of the lignin is achieved by reacting wood chips with a liquor containing sodium hydroxide (NaOH) and sodium sulfide (Na₂S). These compounds typically exist in a 3:1 ratio (as Na₂O) NaOH: Na₂S. Typical pulping conditions reported by Aho are as follows: cooking temperature, 165–175°C; time to achieve maximum temperature, 60–150 min; cooking time at maximum temperature, 60–120 min; liquor:wood ratio, 3–4; and chemical charge, 12–18% active alkali (NaOH + Na₂S, expressed as Na₂O equivalent, is active alkali).

In Kraft pulping the active reagents are HS^- and HO^- . The Na_2S exists in equilibrium with H_2O and serves not only as a source of HS^- , but also as an additional source of NaOH according to the following:



The actual mechanisms of Kraft delignification are highly complex, revolving around the ionization of acid phenolic units in lignin by OH^- and nucleophilic displacement of lignin units with HS^- . The chemistry of delignification is reviewed in detail elsewhere. It is sufficient to note here the conditions specified above and the pulp yields; typically 45–55% of the dry weight of wood furnish is produced as Kraft pulp.

2. Sulfite and Soda Pulping

It is useful to compare Kraft pulping to sulfite pulping as a means for understanding differences among these systems. Such a comparison is shown in Table III. Conditions and results for soda pulping are also shown in Table III. Kraft pulping is presented to facilitate comparison.

From Table III, the similarities and differences among processes become apparent. Certainly the domination of sodium as a base, and sulfur as an active reagent, become obvious. The narrow range of cooking temperatures and yields also becomes apparent. What is not shown is the strength advantage of Kraft pulp. Also not shown are such process considerations as chemical and energy recovery.

3. Other Options

There are numerous alternatives that have been proposed and that are being implemented. These include the addition of anthraquinone to soda and Kraft processes; the use of ferric oxide in the soda pulping process (DARS process) and the substitution of sodium metaborate (NaBO_2) for NaOH in Kraft pulping (borate based autoausticizing). These options largely are designed to achieve process advantages. Anthraquinone (AQ) addition improves pulping yield by 1–3%. Its utility, however, is limited to alkaline systems and its economics are dependent upon the trade-off between raw material and chemical costs. DARS and borate-based Kraft pulping are designed to simplify chemical recovery. The DARS process is applicable only to sulfur-free systems.

Other options include oxygen pulping as well as oxygen bleaching, discussed later in the chapter. A considerable amount of research has been expended on totally new approaches to pulping wood and agro-based materials and include sulfur free organosolv (organic solvent) pulping and biopulping. Organosolv pulping typically employs aqueous organic solvents such as ethanol, methanol or acetic acid as the pulping liquor. Pollution problems are considerably reduced with these methods because the solvents have to be completely recovered for economic reasons; and consequently, this also results in recovery and usage of all the formerly discarded wood components. Another advantage is the potential for developing small, competitive pulp mills with lower capital investment. Two organosolv pulp mills, one each based on 50%

TABLE III Pulping Conditions and Results for Sulfite and Soda Pulping^a

Pulping method	pH range	"Base" alternatives	Active reagents	Max temp (°C)	Time at max temp (min)	Yield (percent)
Acid bisulfite	1–2	Ca^{2+} , Mg^{2+} , Na^+ , NH_4^+	HSO_3^- , H^+	125–145	180–420	45–55
Bisulfite	3–5	Mg^{2+} , Na^+ , NH_4^+	HSO_3^- , H^+	150–170	60–180	50–65
Two-stage sulfite (Stora type)						50–60
Stage 1	6–8	Na^+	HSO_3^- , SO_3^{2-}	135–145	120–360	
Stage 2	1–2	Na^+	HSO_3 , H^+	125–140	120–240	
Three-stage sulfite (Silvola type)						34–45
Stage 1	6–8	Na^+	HSO_3^- , SO_3^{2-}	120–140	120–180	
Stage 2	1–2	Na^+	HSO_3 , H^+	135–145	180–300	
Stage 3	6–10	Na^+	HO^-	160–180	120–180	
NSSC	5–7	Na^+ , NH_4^+	HSO_3^- , SO_3^{2-}	160–180	15–180	75–90 ^b
Alkaline sulfite	9–13	Na^+	SO_3^{2-} , HO^-	160–180	180–300	45–60
Soda	13–14	Na^+	HO^-	155–175	120–300	50–70 ^b
(Kraft)	13–14	Na^+	HO^- , HS^-	155–175	60–18	45–55

^aFrom Sjöström (1981). "Wood Chemistry: Fundamentals and Applications," Academic Press, New York.

^bHardwood.

aqueous ethanol and 85% aqueous acetic acid, were established in Germany in the early 1990s, however, neither mill was successful and both were shut down for a variety of reasons.

With pulping, the inter-fiber lignin bond is broken down by mechanical and/or chemical treatments to free the cellulose fibers for papermaking. In the forest, white rot fungi perform a similar task on wood left behind. The enzymes of the fungi do the work of lignin degradation. This is the basis of new biopulping approaches that have been under development for over 10 years. Wood chips or agricultural materials are treated with a white rot fungus and nutrients for about two weeks which breaks down and alters the lignin gluing substance in the lignocellulosic material. The biomass then can be much more easily disintegrated by mechanical treatment in a disk refiner. Since some mechanical treatment is required the method is more properly termed biomechanical pulping. Investigators at the U.S. Department of Agriculture, Forest Products Laboratory in Madison, WI, evaluated hundreds of fungi for this purpose and found that treatment with the white rot fungus, *Ceriporiopsis subvermispota*, resulted in the greatest reduction in energy requirements for mechanical disintegration and the best strength properties from the resulting paper. Pilot level trails with biomechanical pulping have demonstrated the viability of the process, which is nearing commercial application. All of these new approaches have been reviewed by [Young and Akhtar \(1998\)](#).

C. Process Considerations

Chemical pulping, as performed in the Kraft process, is essentially a closed process. Wood in log form is debarked and chipped. Pulp chips are screened and then sent to continuous or batch digesters. Cooking occurs in the digester where the wood reacts with pulping (white) liquor containing NaOH and Na₂S at elevated temperatures and pressures; following cooking, the chips are “blown” to produce fibers, washed to achieve pulp-liquor separation, and then transported as pulp either to the bleach plant or pulp dryer. The spent pulping (black) liquor is passed through evaporators and concentrators until its moisture content is reduced to about 40%. The black liquor, a mixture of pulping chemicals and dissolved lignin, is then burned in the recovery boiler to achieve energy and chemical recovery. Energy is recovered as high-pressure steam. Chemical recovery is accomplished with sodium carbonate (Na₂CO₃) and sodium sulfide (Na₂S) being tapped from the bottom of the boiler. The smelt is dissolved in water, reacted with calcium oxide from the lime kiln to convert Na₂CO₃ to NaOH, and then returned to the white

liquor. This process is summarized in [Fig. 2](#), a flowsheet of Kraft pulping.

The digester is the heart of the Kraft mill. It may be a continuous digester, such as the unit at Leaf River, Mississippi, shown in [Fig. 3](#). Alternatively batch digesters may be used. The continuous digester offers somewhat higher yields and reduced energy requirements than the batch digester. However, the batch digester offers greater product flexibility.

Kraft pulping requires the consumption of 14–20 GJ/tonne of pulp in the form of heat energy (3–10 atm steam); and 900–1000 kW h/tonne of pulp either as electricity or shaft power. Variation results as much from local economic conditions as from severity of pulping conditions associated with product requirements. The unbleached Kraft pulp mill can generate virtually all of its energy internally with the exception of the 2 GJ/tonne required as oil or gas for the lime kiln. Even there progress is being made in commercializing wood-fired lime kilns.

Yields of 50% and reduced energy consumption have been achieved by a history of innovation. Such innovation has included the Tomlinson black liquor recovery boiler, the Kamyr continuous digester and associated diffusion washer, multiple-effect evaporators, and low-odor concentrators. Economic advantages also have been gained by the development of systems for recovering extractives such as tall oil, fatty acids, and resin from the pulping liquor for sale as naval stores. Future innovations may focus on the lime kiln and other related systems.

Chemical pulping systems other than the Kraft process described earlier also have, at their center, the digester and the recovery system. The major process differences between the Kraft and sulfite pulping methods, from a process perspective, are in the chemical and energy recovery area. [Aho \(1983\)](#) has pointed out that sodium-based systems require highly complex recovery systems such as the Tampella Recovery Process, the Stora Process, the CE Silvola process, and the SCA-Billenid Process. Magnesium based systems permit both energy and chemical recovery; however calcium-based liquor incineration results in a loss of base, a loss of sulfur, and serious scaling problems.

Ammonia-based liquors, when incinerated, result in a loss of nitrogen as N₂ in the flue gas. This difficulty is highly responsible for the domination of Kraft pulping. Magnesium-based liquor incineration is most easily accomplished, and can be achieved either in a Tomlinson furnace or a Copeland fluidized bed system.

While sulfite pulping is less popular than Kraft pulping, it is more prevalent in the production of dissolving pulps. Further, sulfite pulping permits recovery of ethanol from the spent pulping liquor before incineration, as is

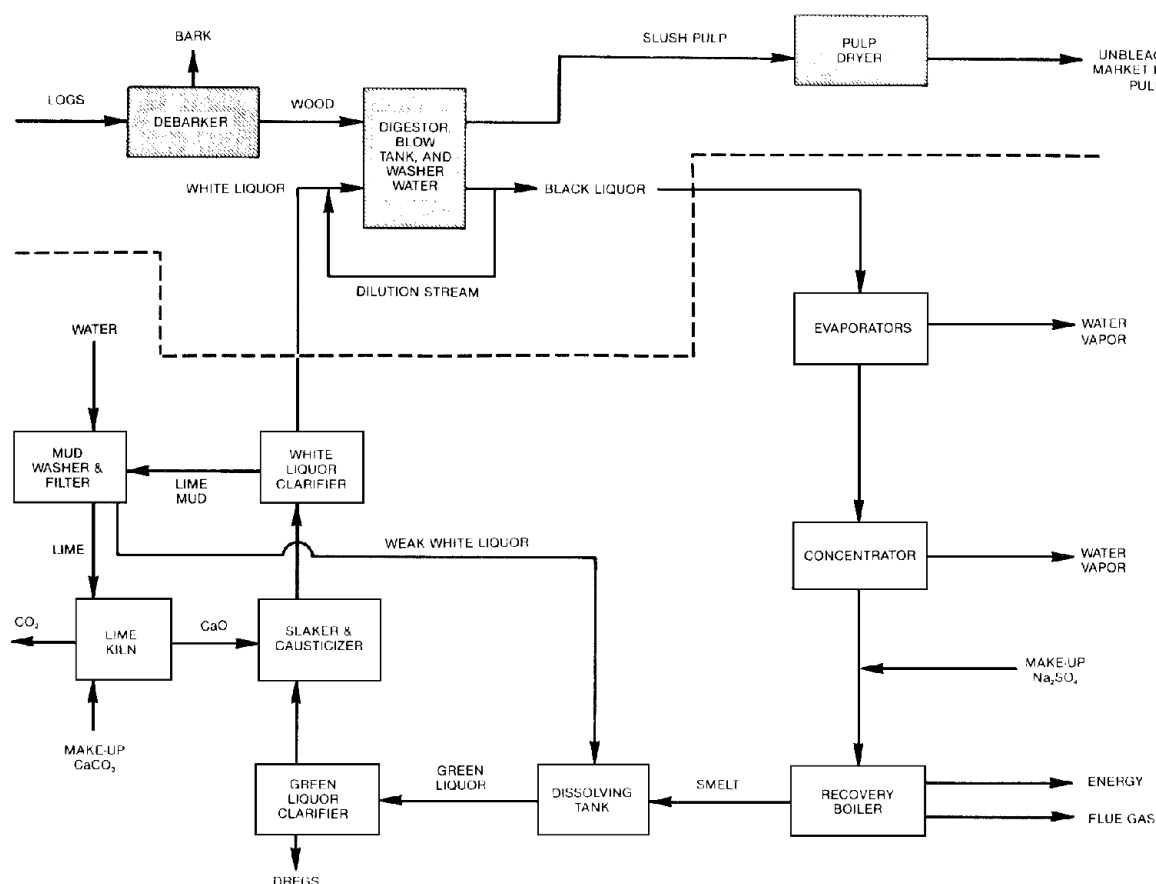


FIGURE 2 Flowsheet of an unbleached Kraft pulp mill focusing on chemical flows. [Reprinted with permission from Tillman, D. A. (1985). "Forest Products: Advanced Technologies and Economic Analysis," Academic Press, Orlando, FL. Copyright 1985 Academic Press.]

performed by the Georgia-Pacific Mill in Bellingham, Washington.

The future of chemical pulping involves process improvements in such areas as liquor recovery, causticizing, and yield improvement. Perhaps more important, however, is the integration of chemical and mechanical pulping as is discussed in the following section.

IV. MECHANICAL PULPING OF WOOD

Industrial pulping processes employ both chemical and mechanical treatment of plant material to provide fiber furnish for subsequent papermaking operations. The proportion of energy applied either chemically or mechanically varies considerably depending upon the desired properties required for a given type of paper. Mechanical pulping is designed for product fibers with certain inherent properties and for taking advantage of the high yields that result from primarily using mechanical energy to fiberize material.

Mechanical pulping was once regarded as describing processes in which yields averaged ~95%. Today the differences between chemical and mechanical pulping are tending to become less apparent. Mechanical pulping now refers to those processes that rely mainly on mechanical means to defiber material.

A. The Range of Mechanical Pulping

Some of the latest process developments in the pulp and paper industry have occurred in the area now broadly defined as mechanical pulping. These processes also represent one of the fastest growth segments in terms of both number and pulp output tonnages. This growth has been accompanied by increasing complexity in the nomenclature describing mechanical pulping processes. Up until 1968, there were basically two types of mechanical pulping techniques: (1) stone ground wood (SGW), and (2) refiner mechanical pulp (RMP).

Stone ground wood pulp, the oldest of the purely mechanical methods, was developed in 1845 and used



FIGURE 3 The modern Kamye continuous digester, heart of the Kraft pulp mill. [Photo courtesy of Leaf River Corp.]

commercially in the 1850s. It still accounts for almost one-half of the total of all mechanical pulp produced worldwide. (25×10^6 tonnes/year). In this process, short bolts of solid wood are pressed against the outer rim of a revolving stone wheel.

Refiner Mechanical Pulp was developed in 1929 and then used in 1938 for board products. Disk refiners began to be used in 1962 for pulp production. In this case, unlike SGW, small wood pieces or chips are broken down between rotating, grooved, or patterned metal disks at atmospheric pressure. The two methods for mechanically producing pulp are depicted in Fig. 4. One advantage of using refiners is that lowercost wood residues could be used as feedstock. Refiner mechanical pulp production totals about 3.5×10^6 tonnes/year.

These first two mechanical methods of breaking down wood into pulp provide the bases for all of the further developments in the field of mechanical pulping. While all of the present mechanical pulping methods do produce different types of pulp, they still rely upon either

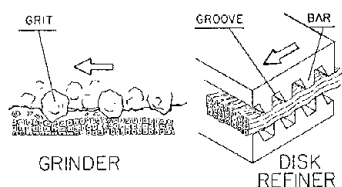


FIGURE 4 A graphic depiction of the various types of mechanical pulping and the relation of grinding systems to fiber dimension.

stone grinding or disk refining to provide the energy for attrition.

Thermomechanical pulp (TMP) was the next step in the development of the newer methods. Thermomechanical pulp, commercially introduced in 1968, was originally designed to reduce the mechanical energy demands of mechanical pulping, but this objective was not achieved. The pulp produced, however, was much stronger than SGW pulp. The SGW and RMP pulps as the sole furnish for paper, are too weak to be used on modern high-speed paper machines. Therefore, up to 25% of high-cost full chemical (e.g., bleached Kraft) fiber is used to reinforce the sheet. There was a need for a process in which the high yields of SGW or RMP could be realized that produce higher quality fiber. In TMP, chips are preheated and converted to pulp in either pressurized or unpressurized disk refiners. Preheating the chips softens the lignin and reduces the fragmentation of wood to produce more whole fiber. The pulp is much stronger, due to the mechanisms shown in Fig. 5. Current production of TMP pulp is about 10×10^6 tonnes/year. Since 1970, developments in the TMP methods as well as other factors have spurred the recent growth in the complexity of mechanical pulping processes as shown in Fig. 6. Since the early 1970s, the number of different mechanical pulping methods has expanded dramatically. There is probably no other period in history in which so many different forms of pulp processing techniques have been developed.

In Fig. 6 all of the methods are divided into purely mechanical pulps and chemically modified pulps. Under purely mechanical pulps the older methods, SGW, RMP, and TMP remain, but three new processes have been added to the list: TRMP (thermo-refiner mechanical pulp), PGW (pressure ground wood) and PRMP (pressure refiner mechanical pulp). These purely mechanical methods are all very similar to the older processes. The differences are related to the temperature of either the wood before or during refining. Heat energy or pressure is not applied in the same manner in the different processes.

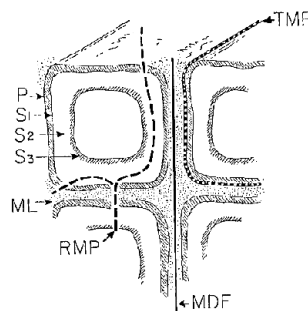


FIGURE 5 The cleavage mechanism of TMP pulping compared to RMP pulping and medium density fiberboard (MDF) production.

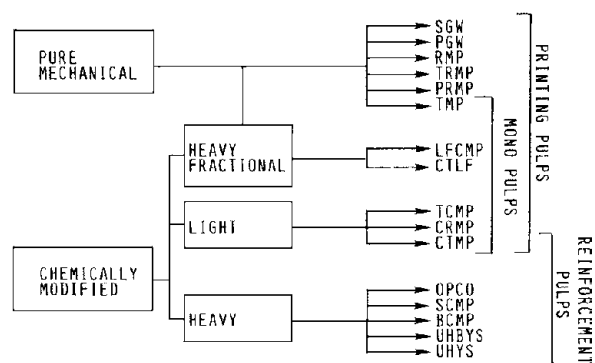


FIGURE 6 The current family of mechanical and chemically modified mechanical pulps.

In the case of PGW, the casing surrounding the pulpstone is sealed and pressurized. This can be accomplished with steam, but most manufacturers now use air pressure. This pressure maintains higher temperatures in the grinding zone, and the yield of longer fiber pulp is increased. PRMP uses air or steam pressure applied to the refiner. The chips are unheated and untreated. TRMP is closely related to PRMP, the chips are preheated and refined at atmospheric pressure.

In summary, purely mechanical pulping methods give the highest yields—93–99%. Advances have been made by modifying older processes by application of heat energy to assist in defiberization. Pure mechanical pulps may have excellent printing properties and optical properties. But, most processes yield fiber that is still too weak to be used without reinforcement—the only exception being TMP and closely related methods. Full TMP furnish newsprints are being made.

Chemically modified pulps, as the name implies, are pulps produced by either subjecting wood chips to a mild chemical treatment or using chemical treatment at some point during or after refining. Often steam is injected with the chemicals to yield a chemithermomechanical pulp (CTMP) with good strength properties. This approach has been adopted by many Canadian mills in recent years.

The chemical pretreatment utilizes chemicals also common to the full chemical pulping processes. However, the chemical treatments are much shorter in duration and generally lower temperatures are used to minimize solubilization of wood components and keep the yields high. Most of these processes use sodium sulfite or bisulfite as the active chemical, although sodium hydroxide, sulfide, and carbonate are also being used. With chemically modified pulps, organization is achieved by dividing these processes into three groups—heavy fractional and light and heavy chemical treatment. The adjectives light and heavy describe the degree of sulfonation applied to the woodchips or fiberized wood.

Several factors are responsible for the incorporation of chemical treatments in mechanical pulping including the high-energy demand of using only mechanical attrition and the limited utility of mechanically pulping many hardwood species.

Mild chemical treatments are used to soften the wood chips and increase the amount of whole fibers. Some of the wood cell components are solubilized and the lignin is made more hydrophilic. The penalty paid for chemical addition is a reduction in yield, to levels of 80 to 90%. The benefits of chemical addition pulps are the increased ability to utilize hardwoods, lower energy requirements, stronger pulps, and increased process flexibility. Some of these pulps have been found to be suitable for products that generally require full chemical pulps.

B. Process Considerations

The modern process of mechanical pulping is best understood in terms of the TMP process. The basic process is depicted, schematically, in Fig. 7. From Fig. 7 it is apparent that the heart of the system is the refiner and TMP systems may have from one to three stages of refiners such as the Sprout–Waldron machine depicted in Fig. 8. When chemical addition is performed, it is in conjunction with chips steaming (see Fig. 7).

The TMP systems can consume 2200–2800 kW h/tonne of pulp, depending upon the species being pulped and the degree of refining employed. Much of the energy consumed ends up as waste heat. Consequently, waste heat recovery is of primary economic importance. Systems such as that depicted in Fig. 9 are used to produce steam for use in the TMP process, and for steam and hot water useful in other forest industry processes. Waste heat recovery has improved the economics of TMP and related mechanical pulping systems, particularly in integrated forest industry mill settings.

C. Prospectus

Mechanical pulping has higher yields but lower-strength pulps when compared to full chemical pulps. Improvements are constantly being made, and considerable gains have been made in adapting different types of wood and different forms of wood (sawdust versus chips) to mechanical pulping via advanced process control techniques.

The pulp and paper industry is undergoing some relatively rapid changes in pulping technology. In areas of the world where the resource base is dwindling, the increased yields offered by newer mechanical pulping techniques are highly desirable. This has been the case in Sweden, Finland, and Canada, which have low-cost hydroelectric power available in many sections.

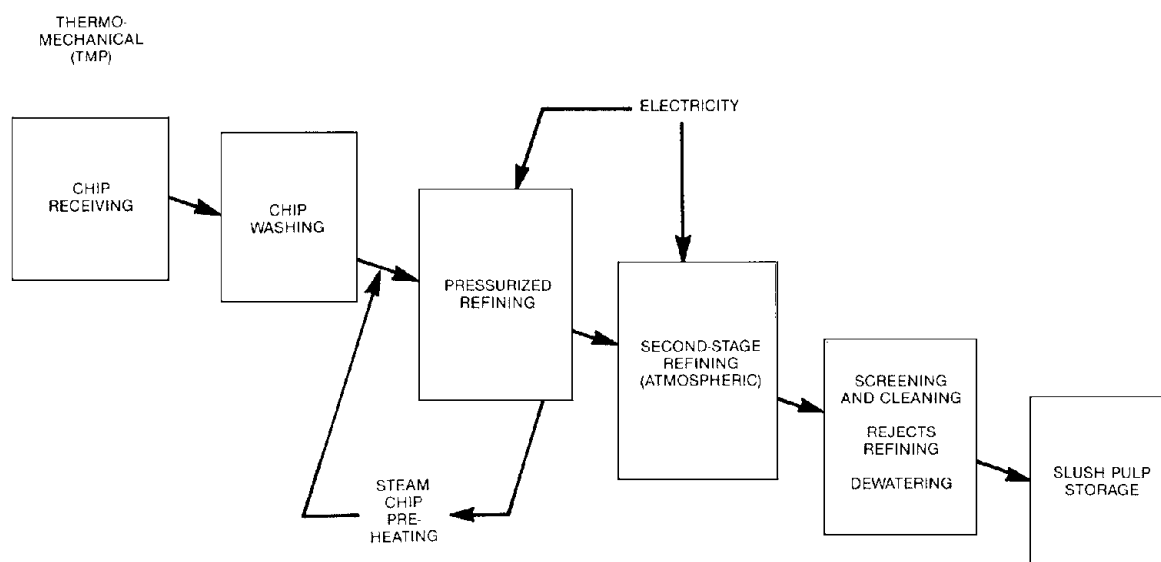


FIGURE 7 Simplified flowsheet of a TMP mill.

V. BLEACHING OF WOOD PULPS

Some wood pulps are used without bleaching for certain paper grades. However, many end uses of paper require further purifying or brightening of the fiber furnish. The color of wood pulp is usually due to the lignin remaining in the fibers after pulping. In the lignin molecule, conjugated single and double bond structures are the primary light-absorbing groups (chromophores) responsible for the color in pulp. Brightness can be increased in two basic ways: color can be removed by either removing the lignin or altering the conjugate double bond structure.

In chemical pulping most of the lignin is already removed. Thus, bleaching is usually accomplished by extracting the remaining lignin. In semichemical or mechanical pulping, very little of the lignin is ever removed. These pulps are generally bleached by chemically altering

the existing chromophoric bond structure to shift light absorption out of the visible range.

A. Chemistry of Bleaching

Presently, three general types of chemicals are used in the bleaching of wood pulp: (1) chlorine containing agents, (2) oxidizing agents, and (3) reducing agents. Chlorine is a major wood pulp bleach. Molecular chlorine can react with lignin by either addition, substitution, or oxidation. The lignin is primarily chlorinated or oxidized. The chlorinated and oxidized lignin is much more soluble than the original lignin and can be subsequently extracted efficiently by caustic solutions. In theory, chlorination could provide all the bleaching power. It is advantageous, however, to use only a portion (60–70%) of the total chlorine demand of a pulp in the initial bleaching step. Subsequent caustic extraction removes 50–90% of the lignin depending on the pulp. These steps also appear to alter the morphological structure and allow milder reactants to modify the remaining lignin more effectively. Chlorine also oxidizes carbohydrates so conditions are selected to optimize the lignin removal. Chlorinations are typically run at a pH of 2–4, at low temperatures and concentrations (consistencies of 3–4%). Residence times are typically under an hour. Chlorine concentrations can vary from 3 to 8%.

Of the oxidative agents, chlorine dioxide (ClO_2) is one of the most effective used to brighten wood pulp. ClO_2 chlorine dioxide is highly specific and bleaches almost any type of pulp, other than high-yield mechanical pulps, to high brightness levels without significant effect on pulp

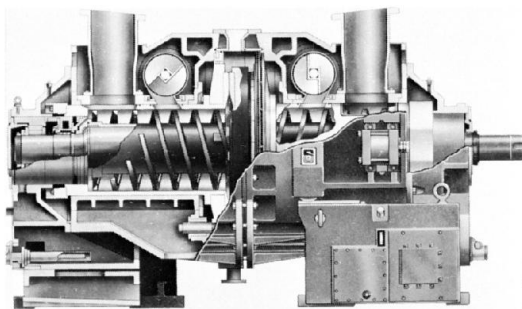


FIGURE 8 The Sprout Waldron Twin Refinery used in TMP Pulping. [Photo courtesy of Sprout-Waldron Co.]

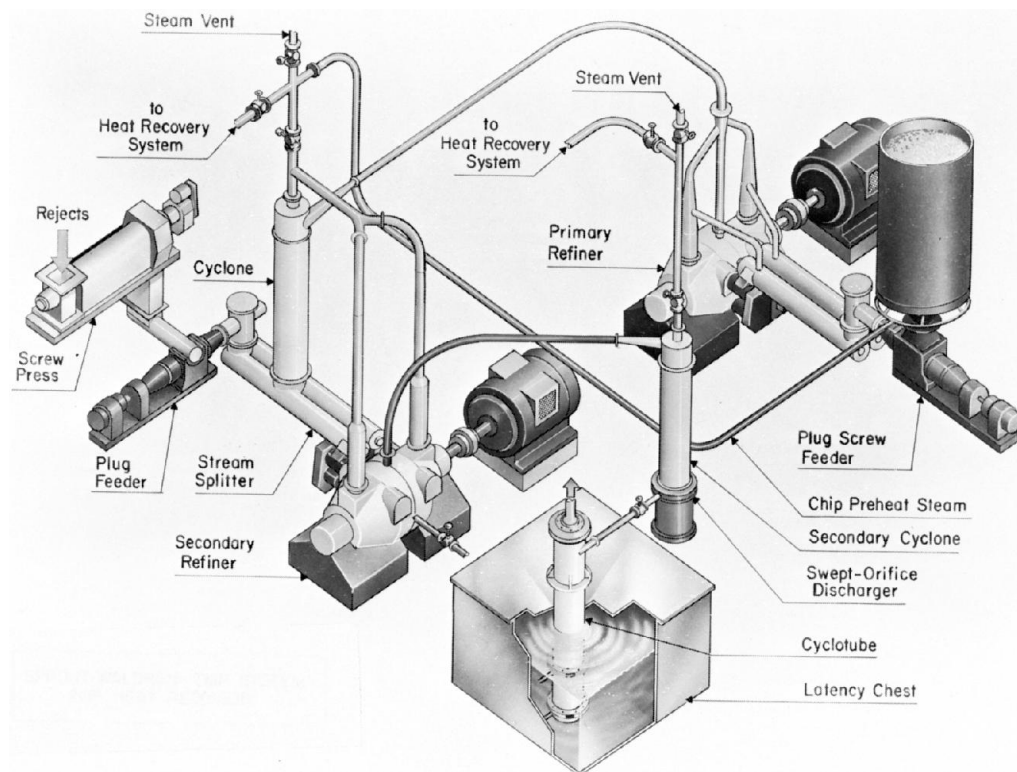


FIGURE 9 The Sprout-Waldron waste heat recovery system used in TMP pulping. [Photo courtesy of Sprout-Waldron Co.]

properties. ClO_2 can completely replace chlorination in multistage bleaching, but because of economics, it is generally used as part of the initial step or in later bleaching sequences. ClO_2 is used in concentrations of 0.5 to 3% at temperatures of 40–50°C and consistencies of 12%. The reactions are held for 2–4 hr.

Chlorination or chlorine dioxide treatment is almost always followed by a caustic extraction. Alkaline extraction neutralizes the pulp and removes the lignin rendered more soluble by chlorine treatment. Removal of this modified lignin opens up the cell wall and allows for milder delignification to be used in later bleaching steps. Caustic extraction is generally performed with most chlorinated chemical pulps at 50–60° with 0.5–5% NaOH for 1 to 2 hr. Consistencies are usually quite high (12 to 16%) to reduce the amount of water and minimize energy requirements.

Prior to the availability of the commercial quantities of chlorine, hypochlorites were the primary chlorine-containing chemicals used in the bleaching of pulp. Some of the easier bleaching pulps (e.g., sulfite) could be bleached to acceptable levels of brightness in one stage. Hypochlorite is a nonspecific oxidizing agent; therefore, its use alone could not brighten most pulps to very high levels without seriously degrading the carbohydrate fraction of the pulp.

Calcium and sodium hypochlorite are both used in the bleaching of wood pulp, primarily as a step in multistage bleaching. The conditions used are a pH of 10–11, temperatures around 100°C for 2–3 hr. Average chlorine use is about 1.5% based on the pulp.

Hydrogen peroxide (H_2O_2) has great utility in bleaching pulp. It enhances brightness in full chemical pulps after multi-stage chlorine-based bleaching steps are completed. This is performed at an alkaline pH, at slightly elevated temperatures (100°C), and with high pulp consistencies. Reaction times are 2–3 hr with peroxide concentrations of 1–3%.

Oxygen has been one of the most highly investigated bleaching chemicals in the last 2 decades. It can be the lowest-cost oxidizing agent available. However, it is a nonspecific bleach and special steps must be taken to protect the carbohydrates from attack. The pulp is usually acid washed to remove heavy metal contaminants such as iron and manganese, then treated with a magnesium salt to limit carbohydrate damage. Oxygen bleaching is best carried out at high consistencies at a highly alkaline pH. Oxygen bleaching is usually termed the oxygen-alkali stage since both oxygen and sodium hydroxide are often used in equivalent amounts. The oxygen-alkali stage could occur as part of an alkaline

extraction step or as an initial bleaching stage. Delignification is generally held to about one half of the lignin present. Oxygen and peroxide bleaching have many commonalities because they both react with organic compounds in similar ways. However, oxygen is used for delignification whereas peroxide eliminates color without lignin removal.

In semichemical or mechanical pulps, the aims of bleaching are to brighten the pulp while retaining as much of the yield as possible. Therefore, the bleaching chemicals used on such pulps are those that alter, but do not remove, the light absorbing molecules. This is accomplished with oxidative or reductive reagents. The predominant oxidative agents are sodium and hydrogen peroxide. Reducing agents include zinc and sodium hydrosulfite (dithionites), sulfur dioxide, sodium sulfite, and bisulfite and sodium borohydride. Of these, hydrogen peroxide and sodium dithionite, represent the greatest use.

The important parameters in bleaching wood pulp are the concentration (consistency) of pulp and bleaching chemicals, the reaction temperature and duration (residence time), the mixing of pulp and chemical, and the pH at which the reactions are carried out. Initial temperatures and concentrations are usually selected based upon experience with a particular pulps needs. Control is achieved by carefully balancing all these various factors to optimize bleaching with a minimum expenditure of chemical.

The bleaching of pulp is done in a carefully balanced series or sequences of treatments. The number and type of bleaching steps is governed by the type of pulp and the end-use requirements. The various bleaching agents discussed above are used to remove or alter the residual lignin in the pulp. These steps are given letter designations by the pulp and paper industry. These designations are chlorine (C), chlorine dioxide (D), hypochlorite (H), peroxide (P) and oxygen (O). A caustic extraction step (E) is usually used at some point between some of the bleaching sequences. If more than one type of bleaching chemical is used in any one step, the minor agent is usually subscripted (i.e., E_0 for caustic extraction with oxygen). If the agents are used in equivalent amounts such as chlorine and chlorine dioxide, the step may be designated as C/D or C + D.

Some classes of pulp such as the sulfites or bisulfites are relatively easy to bleach. These pulps can be bleached with as few as three to five steps (e.g., CEH, CEHEH). Kraft softwood pulps are difficult to bleach to high brightness. From five to seven steps may be required (e.g., CEHEH, CEHDEDP). Kraft hardwood pulps generally are regarded as intermediate in difficulty.

In recent years, the bleaching process has come under some scrutiny because of the potential to form trace quan-

ties of chlorinated dibenzo-dioxins and dibenzofurans, particularly when chlorine is used as the bleaching agent. The pulp and paper industry has demonstrated that substitution of chlorine dioxide for chlorine in the bleaching process significantly reduces if not eliminates the potential for formation of such chlorinated dioxins and furans and the consequent emission of trace quantities of such compounds in pulp mill effluents.

With the removal of chlorine from the bleaching sequence, the process is termed Elemental Chlorine Free (ECF) bleaching and usually an Oxygen (O) stage is now substituted for the Chlorine (C) stage. Regulatory agencies in Europe, and particularly in Scandinavia, have imposed even greater restrictions on emissions from pulp mill bleach plants and another new approach has been developed, namely, Totally Chlorine Free (TCF) bleaching of pulps. For TCF more radical changes are necessary with substitution of both (C) and (D) stages with ozone (O), peroxide (P), and enzyme (X) stages in a sequence such as OXZP.

The use of enzymes is the newest development in bleaching technology. At least one enzyme based process developed in Finland has been applied commercially. The process uses xylanase to make lignin more vulnerable to oxidation by attacking the surrounding polysaccharides that protect the lignin. Another exciting application would be to use of these and other enzymes for removal of lignin pollutants from waste effluents. Biotechnology should lead to safer and cleaner methods for pulping and bleaching.

Bleaching remains an energy intensive and costly part of pulp and paper production. Studies are continuing on reducing the investment in the large facilities required and the water and energy usage. The bleach plant of the future will consist of fewer stages to achieve the brightness levels required of paper.

VI. PAPERMAKING

Paper is a thin sheet of material which, under low-power magnification, appears as a network of very small fibers. These fibers are generally much greater in length than in diameter, and this length to width difference is an important factor in controlling sheet properties. In engineering terms, paper is an orthotropic material (i.e., the mechanical and physical properties of paper vary in each principal, orthogonal direction). When the fibers are first deposited to make the sheet, the fibers are rarely oriented in a completely random manner. Instead, the long axis of the fiber is frequently biased in the direction of machine travel. Thus, paper is stronger in tension, but tears more easily in the machine direction. Since most fibers

swell and shrink more in width than in length, paper is usually more dimensionally stable in the principle fiber direction than in the cross-fiber direction. Other materials may be added to paper in order to improve a particular property.

A. Fiber Preparation

While many factors are important in determining the properties of paper, interfiber bonding is the most significant factor controlling strength of the sheet. The surface of cellulose fibers is very active and is capable of forming secondary bonds (hydrogen bonds) with adjacent cellulose fibers, provided that the surfaces can be brought into very close contact. In paper, the driving force that brings fibers into this close contact is the surface tension created as the water is removed during drying. As fiber flexibility increases, more surface can conform to the adjacent fiber and a higher level of interfiber bonding can occur. The nature of surface bonding is also affected by the chemical makeup. Fiber surfaces high in lignin content do not bond as well as surfaces high in the amount of noncellulosic carbohydrates or hemicelluloses.

After mechanical or, to a lesser extent, chemical pulping, almost all fiber is subjected to some additional degree of mechanical action that is called synonymously either refining or beating. This mechanical action is important for developing strength in paper by increasing interfiber bonding. Chemical pulps are lower in lignin content than mechanical pulp, so refining action can more easily disrupt the internal cell wall material of chemical pulp. Fibrillation is another method of increasing fiber bonding by increasing the surface area of bonding. Following such stock preparation, the fibers are converted into paper.

B. The Paper Machine

Most paper today is made in continuous sheets on high-speed cylinder or fourdrinier machines. In the cylinder machine, a wire-covered cylinder is partially submerged in a slurry of fibers. The fibers are picked up by the wire as the cylinder revolves. The web is then removed at the top of the cylinder and passed into a press section. Cylinder machines are generally made up of a series of cylinders that join additional plies to the forming sheet. Most paperboard is made on cylinder machines. Fourdrinier machines operate by depositing a slurry of fibers onto a moving wire. The wire is supported during travel by a number of devices that aid in water removal before the web passes into the press section. Fourdriniers are the dominant papermaking machines today. They are used for most paper grades from tissues to writing papers.

A third type of paper machine is also utilized to a lesser extent: the twin wire machine. Instead of depositing a fiber slurry onto a moving wire, the fiber dispersions are delivered into the gap of two moving wires. Machines of this type remove water from both top and bottom surfaces by pressure. Twin wire machines are capable of very high speeds.

High-speed paper machines are the result of a balance of the science of engineering and practical empirical observation. In the past, the art often preceded the science, but as machine speeds increase, visual observation of the phenomenon taking place in papermaking is virtually impossible. Today, the thrust in papermaking is toward faster machine speeds while making paper lighter and bulkier, and papermaking is becoming more of a science.

C. The Use of Additives in Papermaking

While paper can be made of wood fibers alone, little is actually made without some chemical addition or modification. These chemical additives are used to either assist in papermaking or to give the paper certain desirable end-use qualities. These chemicals can be added at virtually any step in papermaking. Some of the additives are used to influence the entire sheet properties. These chemicals are added to the pulp slurry prior to sheet formation (internal addition). When the surface properties of the sheet also need to be altered, additives are used on the sheet after some period of formation or drying (external addition). A number of these chemicals serve commonly as both internal or external additions.

Chemicals that aid in the papermaking process can assist by increasing drainage, aid in formation or retention of other additives, or increase wet strength. Other aids are those that reduce undesirable foaming or microbial buildup in the system. Some of these papermaking aids add to the pulp, but others do not and are lost during the papermaking process.

D. Process Considerations in Papermaking

The processes occurring in a high-speed newsprint paper machine have been discussed above. There are several additional considerations of note in the overall process picture. Paper for the most part is a commodity item (i.e., production costs are more economical per unit when large tonnages of uniform specifications are produced). Most mills have a break-even point at an 85% capacity so it is vital to operate mills at design capacity. Economies of scale are also found for pulp and paper mills at levels of about 1000 tonnes of paper per day for full chemical mills and 200–400 tonnes of paper per day for semichemical or mechanical mills. Thus, the outputs of paper mills are

enormous in terms of square meters of product produced per day. Any improvement in yield in any step of manufacture must be matched up or down the process in order to benefit production.

If a papermill produced paper with a basis weight of 65 g/m² (40 lb/3000 ft²) at a level of 1000 tonnes per day, a paper machine 7 m wide would have to run at speeds in excess of 1500 m/min. (almost 60 miles per hour) to produce this output. Most mills of this size would have more than one paper machine. Any upsets in the papermaking process are very costly. Any incremental improvements made in pulping yield must be matched by increasing the paper machine speeds to utilize the extra furnish because paper machines are too expensive to replace.

One of the most energy intensive processes in papermaking is the drying of the paper sheet. Paper starts as a slurry consisting of mostly water (99%). Draining and wet pressing will remove much of this water, but a mill of 1000 tonnes/day capacity must remove large amounts of water by evaporation using heat or mechanical energy.

The 1000 tonnes/day mill producing a web out of the press at a temperature of 45°C with a solids content of 30% will need to supply about 5.5 GJ/tonnes paper to the web to remove the excess water.

E. Types of Paper

The pulp and paper industry in the United States produced almost 100 million metric tons of paper and paperboard in 1999. This production was almost equally split between paper and paperboard products, described below:

1. **Linerboard and Corrugating Medium:** Corrugated board is generally the familiar “cardboard” 3-ply laminate consisting of two paperboard facers (liners) adhered to each side of a fluted core. This construction gives high stiffness and strength with the benefit of low cost and weight. For special uses the construction could be varied to include more plies or just a single face liner. Almost two-thirds of paperboard output goes into producing corrugated box or container products. The linerboard is frequently made of Kraft pulp, which is known for its high strength. The fluted medium is an outlet for semichemical pulp or recycled paper, which is used for stiffness rather than high strength.

2. **Other Paperboard:** Remaining paperboard products consist of familiar products such as board for packaging foodstuffs, shirt board, and tablet backs. Some paperboard is bleached for use in packaging food products. Milk carton stock is an example of the use of bleached board.

3. **Fine and Specialty Papers:** Paper products in this category include writing or business and technical paper.

Bristols or card stock is often thought of belonging in the paperboard category, but it more rightly belongs as a paper product.

The important properties required of these papers are surfaces suitable for printing or writing and dimensional stability. About one-half of all paper produced is used for printing or writing. Today, much printed and information-carrying paper is handled in automated machines; therefore, the dimensional stability or the ability of paper to hold its dimensional tolerance is of prime importance.

Paper packaging such as Kraft paper sacks and wrapping paper, either bleached or natural, comprise about 7% of the total production of paper and paperboard. Competition from plastics is strongly impacting paper for packaging. The familiar brown Kraft paper sack is rapidly being replaced by plastic bags. Computer information storage systems have not yet had the expected impact on the total amount of printing paper being made, and in some cases, the advent of low-cost word processing may have increased the production of certain grades.

4. **Newsprint:** Newsprint is a commodity product with standardized properties. The requirements for newsprint are low cost, printability, and runnability. Printability may seem of obvious importance to the reader, but the concept has many facets. The paper must accept ink without excess penetration or feathering. The paper surface must be strong enough to resist linting or fiber pick, which can seriously affect print quality. Opacity is another important facet in printing. Newsprint is a relatively lightweight, thin paper and is printed on both sides. Opposite side show-through seriously affects readability. The low cost and high opacities found in mechanical pulp make this fiber especially suited for newsprint.

Runnability is often expressed as the number of paper web breaks per 100 rolls. Paper web breaks during press runs are very costly in terms of money and time. Thus, high-speed newsprint presses also require newsprint with good tensile properties. Although the tensions in newsprint presses are kept well below the average tensile strengths of the paper, breaks do occur. These are generally the result of some defect in the paper such as shives that are not removed from the mechanical pulp fraction of newsprint.

5. **Tissue:** Tissue refers to a wide variety of lightweight paper from toilet tissue to napkins, towelling, wrapping, and book tissue. The requirements for tissue are softness or bulk, absorbancy, and strength. The appearance or purity of tissue is also very important, and tissues are generally made of bleached pulp. Some of the chemically modified mechanical pulps are finding increased usage in the tissue market. Many tissue grades require a high degree of absorbancy while maintaining wet strength. Special

additives are being developed that increase wet strength without affecting absorbancy.

VII. RECYCLING IN PULP AND PAPER

Recycling is one of the traditional sources of fiber for the pulp and paper industry, and in recent years, it has become an increasingly important element of fiber supply. Periodic fiber shortages coupled with governmental policies have encouraged the increased utilization of recycling. Currently several grades of fiber are used significantly as secondary fiber, including old corrugated containers (OCC), old newsprint, old magazines, and high-grade deinking. Much of the fiber being recycled comes either from industrial scrap (e.g., trimmings from converting facilities), newspaper and magazine overruns, or selected office wastes.

Postconsumer waste paper is being used increasingly, although such materials as mixed waste paper still have limited market acceptance. In 1975 wastepaper recycling was about 25%, but due to environmental pressures the paper industry now recycles 45% of the stock (47 million metric tonnes). Even higher recycling levels, up to 60%, are possible and fiber poor countries such as the Netherlands and Japan are near this level. The use of recycled fiber is limited by the extent of contamination plus final paper and paperboard product specifications such as tear strength, brightness, and regulatory issues (e.g., paper from secondary fiber cannot come into direct contact with food products).

The processing of secondary fiber typically involves hydropulping, a mechanical pulping process for fiber liberation from waste products. Hydropulpers also provide for removal of large tramp objects through "raggers" and "junkers." After hydropulping, the fibers are cleaned through a series of screens. Deinking processes are then used. The selection of deinking process is dependent upon the secondary fiber being processed and the product being made. Certain waste papers are proving increasingly difficult to deink, particularly office papers from dry copiers and laser printers. Deinking may be followed by secondary fiber bleaching, depending upon the quality of the fiber being processed and the final product characteristics required. Secondary fiber pulping and bleaching concentrates contaminants contained in the waste paper. Typically pulping and bleaching of secondary fiber can generate 400–800 lb of wastewater treatment solids (sludge) per ton of incoming secondary fiber, depending upon the type of fiber accepted and the final product produced. Further, secondary fiber operations generate significant quantities of waste from the ragger and from the primary and secondary screens used to clean the hydropulper product.

These wastes are disposed of either by incineration or land disposal.

Once recycled pulps are produced they are either blended with virgin fiber for use in paper products or used exclusively. Blending affords the opportunity to gain the characteristics of strength and brightness associated with longer fibers from virgin (wood) sources. Typically, for example, repulped newsprint is blended with TMP pulps as a means to produce acceptable feeds for making new newsprint. The TMP pulps provide the long fiber and consequent strength required for high-speed paper machines and high-speed printing equipment. Blending also may take the form of multi-ply sheet forming in the papermaking process. Typical products that have high secondary fiber utilization include newsprint, folding boxboard, corrugating medium, moulded pulp trays, and certain construction papers. Recycling, then, provides an alternative source of fiber to the pulp and paper industry. This source of fiber is used as a consequence of both raw material and governmental pressures. It is used in specific processes and in selected products, depending upon the source of the secondary fiber and the consumer acceptance of the final product with characteristics imparted by utilization of recycled product. Recycling, then, has become an increasingly important element of the pulp and paper industry.

SEE ALSO THE FOLLOWING ARTICLES

BIOPOLYMERS • CARBOHYDRATES • ENERGY FLOWS IN ECOLOGY AND IN THE ECONOMY

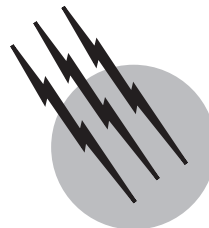
BIBLIOGRAPHY

- Aho, W. (1983). Advances in chemical pulping processes in progress. In "Progress in Biomass Conversion," Vol. 4, Academic Press, New York.
- Biermann, C. J. (1993). "Essentials of Pulping and Papermaking," Academic Press, New York.
- Breck, D. H. (1985). "Technological advances hold the key," *Tappi* **68**(4), 71–72.
- Casey, J. P. (1980). "Pulp and Paper Chemistry and Technology," 3rd Ed., Vol. 1, Wiley (Interscience), New York.
- Fengel, D., and Wegener, G. (1984). "Wood: Chemistry, Ultrastructure and Reactions," Walter de Gruyter Pub., New York.
- Hersch, H. N. (1981). "Energy and Materials Flows in the Production of Pulp and Paper," Argonne National Laboratory, Chicago, IL.
- Libby, C. E. (1962). "Pulp and Paper Science and Technology," Vol. 1, Pulp. McGraw-Hill, New York.
- Mark, R. E. (1983). "Handbook of Physical and Mechanical Testing of Paper and Paperboard," Vols. 1 and 2, Marcel Dekker, New York.
- Tillman, D. A. (1985). "Forest Products: Advanced Technologies and Economic Analyses," Academic Press, Orlando, FL.

- Young, R. A. (1992). Wood and wood products. *In* "Riegel's Handbook of Industrial Chemistry," 9th Ed. (J. Kent, ed.), Van Nostrand Reinhold Pub., New York.
- Young, R. A. (1997). Processing of agro-based resources into pulp and paper. *In* "Paper and Composites from Agro-Based Resources"

(R. Rowell and R. A. Young, eds.), Lewis Pub., CRC Press, Boca Raton, FL.

- Young, R. A., and Akhtar, M. (1998). "Environmentally Friendly Technologies for the Pulp and Paper Industry," John Wiley & Sons Pub., New York.



Reactors in Process Engineering

Gary L. Foutch
Arland H. Johannes

Oklahoma State University

- I. Reactor Classifications
- II. Primary Reactors
- III. Generalized Reactor Design
- IV. Special Reactor Configurations

GLOSSARY

Adiabatic reactor Vessel that is well insulated to minimize heat transfer and has an increase or decrease in temperature from the initial inlet conditions due solely to the heats of reaction.

Batch reactor Vessel used for chemical reaction that has no feed or effluent streams. The reactor is well stirred and usually run either isothermally or adiabatically. The main design variable is how much time the reactants are allowed to remain in the reactor to achieve the desired level of conversion.

Catalyst Substance that increases the rate of a chemical reaction without being consumed in the reaction.

Continuous stirred tank reactor Sometimes called a continuous-flow stirred-tank reactor, ideal mixer, or mixed-flow reactor, all describing reactors with continuous input and output of material. The outlet concentration is assumed to be the same as the concentration at any point in the reactor.

Conversion Fraction or percentage that describes the extent of a chemical reaction. Conversion is calculated by dividing the number of moles of a reactant that reacted

by the initial moles of reactant. Conversion is defined only in terms of a reactant.

Elementary reaction Reaction that has a rate equation that can be written directly from a knowledge of the stoichiometry.

Isothermal reactor Any type of chemical reactor operated at constant temperature.

Mean residence time Average time molecules remain in the reactor. Note that this is different from space time.

Multiple reactions Series or parallel reactions that take place simultaneously in a reactor. For example, $A + B \rightarrow C$ and $A + D \rightarrow E$ are parallel reactions, and $A + B \rightarrow C + D \rightarrow E + F$ are series reactions.

Plug flow reactor Sometimes called a piston flow or a perfect flow reactor. The plug flow reactor has continuous input and output of material. The plug flow assumption generally requires turbulent flow. No radial concentration gradients are assumed.

Product distribution Fraction or percentage of products in the reactor effluent.

Rate constant Constant that allows the proportionality between rate and concentration to be written as a mathematical relationship. The rate constant is a

function of temperature only and is generally modeled by an exponential relationship such as the Arrhenius equation.

Rate equation Mathematical expression that is a function of both concentration of reactants or products, and temperature.

Reaction mechanism Series of elementary reaction steps that when combined, gives the overall rate of reaction.

Space time Time to process one reactor volume based on inlet conditions.

Yield Moles of a desired product divided by moles of a limiting reactant.

A CHEMICAL REACTOR is any type of vessel used in transforming raw materials to desired products. The vessels themselves can be simple mixing tanks or complex flow reactors. In all cases, a reactor must provide enough time for chemical reaction to take place.

The design of chemical reactors encompasses at least three fields of chemical engineering: thermodynamics, kinetics, and heat transfer. For example, if a reaction is run in a typical batch reactor, a simple mixing vessel, what is the maximum conversion expected? This is a thermodynamic question answered with knowledge of chemical equilibrium. Also, we might like to know how long the reaction should proceed to achieve a desired conversion. This is a kinetic question. We must know not only the stoichiometry of the reaction but also the rates of the forward and the reverse reactions. We might also wish to know how much heat must be transferred to or from the reactor to maintain isothermal conditions. This is a heat transfer problem in combination with a thermodynamic problem. We must know whether the reaction is endothermic or exothermic.

After chemical reaction a series of physical treatment steps is usually required to purify the product and perhaps recycle unreacted material back to the reactor. The quantity of material to be processed is a key factor in determining what type of reactor should be used. For small-lot quantities, a batch reactor is commonly used in industry. For large, high-volume reactions, such as in the petroleum industry, flow reactors are common.

I. REACTOR CLASSIFICATIONS

Reactors may be classified by several different methods depending on the variables of interest. There is no single clear cut procedure for reactor classification. As a result, several of the more common classification schemes are presented here.

A. Operation Type

The operational configuration for the reactor can be a primary method of classification.

1. Batch

Batch reactors are operated with all the material placed in the reactor prior to the start of reaction, and all the material is removed after the reaction has been completed. There is no addition or withdrawal of material during the reaction process.

2. Semibatch

The semibatch reactor combines attributes of the batch and the continuous-stirred tank. The reactor is essentially batch but has either a continuous input or output stream during operation.

3. Continuous Flow Reactors

Continuous flow reactors represent the largest group of reactor types by operational classification. Several continuous flow reactors are used industrially.

a. The continuous-stirred tank reactor (CSTR) involves feeding reactants into a well-mixed tank with simultaneous product removal.

b. The plug flow reactor (PFR) consists of a long pipe or tube. The reacting mixture moves down the tube resulting in a change in concentration down the length of the reactor.

c. In the recycle reactor part of the outlet stream is returned to the inlet of the reactor. Although not a typical reactor classification by type, the recycle reactor allows for continuous operation in regimes between CSTR and PFR conditions.

B. Number of Phases

Reactors can also be classified by the number of phases present in the reactor at any time.

1. Homogeneous

Homogeneous reactors contain only one phase throughout the reactor.

2. Heterogeneous

Heterogeneous reactors contain more than one phase. Several heterogeneous reactor types are available due to various combinations of phases.

- a. Gas–liquid
- b. Gas–solid
- c. Liquid–solid
- d. Gas–liquid–solid

Multiphase reactor configurations are strongly influenced by mass transfer operations. Any of the reactor types presented above can be operated as multiphase reactors.

C. Reaction Types

Classification of reactors can also be made by reaction type.

1. Catalytic. Reactions that require the presence of a catalyst to obtain the rate conditions necessary for that particular reactor design
2. Noncatalytic. Reactions that do not include either a homogeneous or heterogeneous catalyst
3. Autocatalytic. Reaction scheme whereby one of the products increases the overall rate of reaction
4. Biological. Reactions that involve living cells (enzymes, bacteria, or yeast), parts of cells, or products from cells required for the reaction scheme
5. Polymerization. Reactions that involve formation of molecular chains, whether on a solid support or in solution.

D. Combination of Terms

Any combination of the previously mentioned classifications can be used to describe a reactor: for example, a heterogeneous-catalytic-batch reactor.

II. PRIMARY REACTORS

There are five primary reactor designs based in theory: batch, semibatch, continuous-stirred tank, plug flow, and fluidized bed. The operating expressions for these reactors are derived from material and energy balances, and each represents a specific mode of operation. Selected reactor configurations are presented in Fig. 1.

A. Batch

1. Description

Batch processes are the easiest to understand since they strongly relate to “cookbook” technology. You put everything in at the beginning and stop the reaction at some time

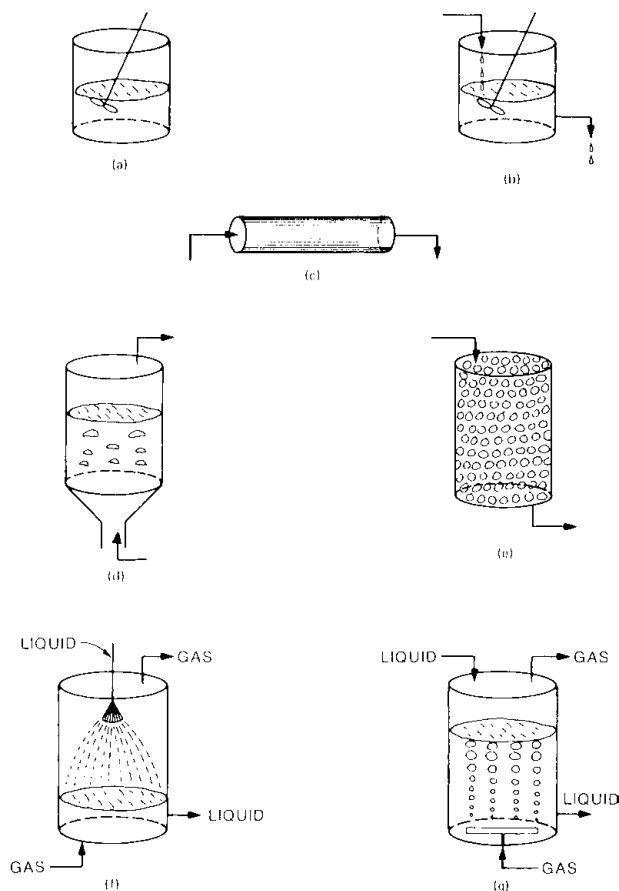


FIGURE 1 Selected reactor configurations: (a) batch, (b) continuous stirred-tank reactor, (c) plug flow reactor, (d) fluidized bed, (e) packed bed, (f) spray column, and (g) bubble column.

later. This cookbook technology allows for immediate production of a new product without extensive knowledge of the reaction kinetics.

The reactor is characterized by no addition of reactant or removal of product during the reaction. Any reaction being carried out with this constraint, regardless of any other reactor characteristic, is considered batch. The assumptions for batch operation are (1) the contents of the tank are well mixed, (2) reaction does not occur to any appreciable degree until filling and startup procedures are complete, and (3) the reaction stops when quenched or emptied. The reactor can be operated with either a homogeneous or heterogeneous reaction mixture for almost any type of reaction.

2. Classification

The batch reactor, one of the five primary reactor configurations, is the oldest reactor scheme.

3. Design Parameters

The design parameters for a batch reactor can be as simple as concentration and time for isothermal systems. The number of parameters increases with each additional complication in the reactor. For example, an additional reactant requires measurement of a second concentration, a second phase adds parameters, and variation of the reaction rate with temperature requires additional descriptors: a frequency factor and an activation energy. These values can be related to the reactor volume by the equations in Section III.

4. Applications

Application of the batch reactor design equations requires integration over time. Along with the simplicity of cookbook chemistry, this is one of the major advantages of the batch reactor: concentrations are not averaged over time. Initially, when concentrations are at their highest, the corresponding rates of reaction are also high. This gives the greatest amount of conversion in the shortest time. The integral reactor design form makes the batch reactor attractive for higher-order reactions. Batch is also good for reactions in series (if the reaction can be quickly quenched), where large amounts of an intermediate can be produced quickly before it has time to react away to a by-product.

The batch reactor is extremely flexible compared with continuous reactor configurations. For example, temperature can easily be made a function of reaction time. Once the reactor is put into service, operational alternatives are still available. The tank can be operated half-full without affecting product quality, or the reaction time can be modified easily. Both of these changes may cause heat and mass transfer problems in fixed-volume continuous equipment. This flexibility is worthwhile for products that are made in various grades, have seasonal demand, or have subjective specifications such as the taste of beer.

Batch reactors are used extensively in industries where only small quantities of product are made, such as pharmaceuticals. For small amounts, the economy of scale hurts flow reactors, which typically have a higher initial investment for controls and plumbing.

5. Advantages–Disadvantages

The primary advantages of the batch reactor are simplicity of design, which allows for tremendous flexibility, and integration of the performance equation over time. The simplicity of design, usually a stirred tank, makes operation and monitoring easy for the majority of reactions. The

integrated form of the performance equation has varied significance depending on the particular reaction scheme being performed. For example, molecular weight distributions in polymerization reactions can be controlled more precisely in batch reactors.

One of the traditional disadvantages of the batch reactor has been the labor required between runs for emptying and filling the tank. With recent advances in computer control, this disadvantage no longer exists. If the advantages of batch are significant, the capital expense of computer control is essentially negligible. Due to computer control, the batch reactor should no longer be looked upon as something to be avoided. If the kinetics and design parameters indicate that batch is a competitive design, then use it.

The major disadvantage of batch reaction now is the hold-up time between batches. Although the actual reaction time necessary to process a given amount of feed may be substantially less than for a time-averaged reactor such as a CSTR, when the hold-up time is added, the total process time may be greater. Other disadvantages of the batch reactor are dependent on the particular type of reaction being considered, such as whether the reaction is in parallel or series.

B. Semibatch

1. Description

The semibatch reactor is a cross between an ordinary batch reactor and a continuous-stirred tank reactor. The reactor has continuous input of reactant through the course of the batch run with no output stream. Another possibility for semibatch operation is continuous withdrawal of product with no addition of reactant. Due to the crossover between the other ideal reactor types, the semibatch uses all of the terms in the general energy and material balances. This results in more complex mathematical expressions. Since the single continuous stream may be either an input or an output, the form of the equations depends upon the particular mode of operation.

Physically, the semibatch reactor looks similar to a batch reactor or a CSTR. Reaction occurs in a stirred tank, with the following assumptions; (1) the contents of the tank are well mixed, and (2) there are no inlet or outlet effects caused by the continuous stream.

2. Classification

The semibatch reactor is one of the primary ideal reactor types since it can not be accurately described as either a continuous or a batch reactor. A semibatch reactor is usually classified as a type of transient reactor.

3. Design Parameters

The major design parameters for a semibatch reactor are similar to a batch reactor with the addition of flow into or out of the tank.

4. Applications

The advantage of this reactor, with feed only, is for the control of heat of extremely exothermic reactions. By inputting the feed gradually during the course of the reaction, the concentration of feed in the reactor can be kept lower than in normal batch operation. Also, the temperature of the feed stream, when cooler than the reaction mixture, has a quenching effect. Some of the heat released during the reaction is used to heat the feed material, thereby reducing the required capacity of the heating coils.

The semibatch can also be used to control the kinetics in multiple reaction sequences. The selectivity may be shifted to one reaction by adding a reactant slowly. This keeps one reactant concentration high with respect to the other.

The semibatch can also be used for continuous product removal, such as vaporization of the primary product. This can increase yield in equilibrium limited reactions.

5. Advantages–Disadvantages

The temperature-controlling features of this reaction scheme dominate selection and use of the reactor. However, the semibatch reactor does have some of the advantages of batch reactors: temperature programming with time and variable reaction time control.

The temperature conditions and the batch nature of this reactor are the primary operational difficulties and make the reactor impractical for most reactions, even for computer-controlled systems. The majority of reactions considered for semibatch are highly exothermic and, as such, are dangerous and require special attention.

C. Continuous-Stirred Tank

1. Description

The continuous-stirred tank reactor (CSTR) has continuous input and output of material. The CSTR is well mixed with no dead zones or bypasses in ideal operation. It may or may not include baffling. The assumptions made for the ideal CSTR are (1) composition and temperature are uniform everywhere in the tank, (2) the effluent composition is the same as that in the tank, and (3) the tank operates at steady state.

We traditionally think of the CSTR as having the appearance of a mixing tank. This need not be the case. The

previously mentioned assumptions can be met even in a long tube if the mixing characteristics indicate high dispersion levels in the reactor. This is particularly true of gassed liquids where the bubbling in the column mixes the liquid.

2. Classification

The continuous-stirred tank reactor is one of the two primary types of ideal flow reactors. It is also referred to as a mixed-flow reactor, back-mix reactor, or constant-flow stirred-tank reactor.

3. Design Parameters

The CSTR is not an integral reactor. Since the same concentration exists everywhere, and the reactor is operating at steady state, there is only one reaction rate at the average concentration in the tank. Since this concentration is low because of the conversion in the tank, the value for the reaction rate is also low. This is particularly significant for higher-order reactions compared with integral reactor systems.

Time is still an important variable for continuous systems, but it is modified to relate to the steady-state conditions that exist in the reactor. This time variable is referred to as space time. Space time is the reactor volume divided by the inlet volumetric flow rate. In other words, it is the time required to process one reactor volume of feed material. Since concentration versus real time remains constant during the course of a CSTR reaction, rate-data acquisition requires dividing the difference in concentration from the inlet to the outlet by the space time for the particular reactor operating conditions.

4. Applications

The CSTR is particularly useful for reaction schemes that require low concentration, such as selectivity between multiple reactions or substrate inhibition in a chemostat (see Section IV). The reactor also has applications for heterogeneous systems where high mixing gives high contact time between phases. Liquid–liquid CSTRs are used for the saponification of fats and for suspension and emulsion polymerizations. Gas–liquid mixers are used for the oxidation of cyclohexane. Gas homogeneous CSTRs are extremely rare.

5. Advantages–Disadvantages

The advantages for CSTRs include (1) steady-state operation; (2) back mixing of heat generated by exothermic reactions, which increases the reaction rate and subsequent

reactor performance; (3) avoidance of reactor hot spots for highly exothermic reactions, making temperature easier to control; (4) favoring lower-order reactions in parallel reaction schemes; (5) economical operation when large volumes require high contact time; and (6) enhancement of heat transfer by mixing.

For the kinetics of decreasing rate with increasing conversion (most reactions), isothermal CSTRs have lower product composition than plug flow reactors. Additional disadvantages of CSTR are that larger reactor volumes are usually required, compared with other reactor schemes, and that energy for agitation is required in the tank, increasing operating costs.

D. Plug Flow

1. Description

This reactor has continuous input and output of material through a tube. Assumptions made for the plug flow reactor (PFR) are (1) material passes through the reactor in incremental slices (each slice is perfectly mixed radially but has no forward or backward mixing between slices; each slice can be envisioned as a miniature CSTR), (2) composition and conversion vary with residence time and can be correlated with reactor volume or reactor length, and (3) the reactor operates at steady state.

The PFR can be imagined as a tube, but not all tubular reactors respond as PFRs. The assumptions need to be verified with experimental data.

2. Classification

The plug flow reactor is the second primary type of ideal flow reactor. It is also erroneously referred to as a tubular reactor.

3. Design Parameters

The parameters for PFRs include space time, concentration, volumetric flow rate, and volume. This reactor follows an integral reaction expression identical to the batch reactor except that space time has been substituted for reaction time. In the plug flow reactor, concentration can be envisioned as having a profile down the reactor. Conversion and concentration can be directly related to the reactor length, which in turn corresponds to reactor volume.

4. Applications

For normal reaction kinetics the plug flow reactor is smaller than the continuous-stirred tank reactor under similar conditions. This gives the PFR an advantage over

CSTR for most reactions. These conditions are best met for short residence times where velocity profiles in the tubes can be maintained in the turbulent flow regime. In an empty tube this requires high flow rates; for packed columns the flow rates need not be as high. Noncatalytic reactions performed in PFRs include high-pressure polymerization of ethylene and naphtha conversion to ethylene. A gas-liquid noncatalytic PFR is used for adipinic nitrile production. A gas-solid PFR is a packed-bed reactor (Section IV). An example of a noncatalytic gas-solid PFR is the converter for steel production. Catalytic PFRs are used for sulfur dioxide combustion and ammonia synthesis.

5. Advantages-Disadvantages

The advantages of a PFR include (1) steady-state operation, (2) minimum back mixing of product so that concentration remains higher than in a CSTR for normal reaction kinetics, (3) minimum reactor volume in comparison with CSTR (since each incremental slice of the reactor looks like an individual CSTR, we can operate at an infinite number of points along the rate curve), (4) application of heat transfer in only those sections of the reactor where it is needed (allowing for temperature profiles to be generated down the reactor), and (5) no requirement for agitation and baffling.

The plug flow reactor is more complex than the continuous-stirred tank alternative with regard to operating conditions. There are a few other disadvantages associated with the PFR. For the kinetics where rate increases with conversion (rare), an isothermal plug flow reactor has lower product composition than a CSTR. For highly viscous reactants, problems can develop due to high-pressure drop through the tubes and unusual flow profiles.

E. Fluidized Bed

1. Description

Fluidization occurs when a fluid is passed upward through a bed of fine solids. At low flow rates the gases or liquids channel around the packed bed of solids, and the bed pressure drop changes linearly with flow rate. At higher flow rates the force of the gas or liquid is sufficient to lift the bed, and a bubbling action is observed. During normal operation of a fluidized bed the solid particles take on the appearance of a boiling fluid. The reactor configuration is usually a vertical column. The fluidized solid may be either a reactant, a catalyst, or an inert. The solid may be considered well mixed, while the fluid passing up through the bed may be either plug flow or well mixed depending on

the flow conditions. Bubble size is critical to the efficiency of a fluidized bed.

2. Classification

Fluidized reactors are the fifth type of primary reactor configuration. There is some debate as to whether or not the fluidized bed deserves distinction into this classification since operation of the bed can be approximated with combined models of the CSTR and the PFR. However, most models developed for fluidized beds have parameters that do not appear in any of the other primary reactor expressions.

3. Design Parameters

In addition to the usual reactor design parameters, height of the fluidized bed is controlled by the gas contact time, solids retention time, bubble size, particle size, and bubble velocity.

4. Applications

Fluidized beds are used for both catalytic and noncatalytic reactions. In the catalytic category, there are fluidized catalytic crackers of petroleum, acrylonitrile production from propylene and ammonia, and the chlorination of olefins to alkyl chlorides. Noncatalytic reactions include fluidized combustion of coal and calcination of lime.

5. Advantages–Disadvantages

The fluidized bed allows for even heat distribution throughout the bed, thereby reducing the hot spots that can be observed in fixed-bed reactors. The small particle sizes used in the bed allow high surface area per unit mass for improved heat and mass transfer characteristics. The fluidized configuration of the bed allows catalyst removal for regeneration without disturbing the operation of the bed. This is particularly advantageous for a catalyst that requires frequent regeneration.

Several disadvantages are associated with the fluidized bed. The equipment tends to be large, gas velocities must be controlled to reduce particle blowout, deterioration of the equipment by abrasion occurs, and improper bed operation with large bubble sizes can drastically reduce conversion.

III. GENERALIZED REACTOR DESIGN

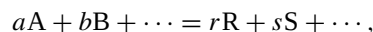
Design of a chemical reactor starts with a knowledge of the chemical reactions that take place in the reactor. The

ultimate product of the design is the reactor and the supporting equipment such as piping, valves, control systems, heat exchangers, and mixers. The reactor must have sufficient volume to handle the capacity required and to allow time for the reaction to reach a predetermined level of conversion or yield.

A. Approach, Considerations, and Methods

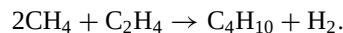
1. Use of the Reaction Coordinate or Molar Extent of Reaction

In chemical reactor design, an understanding of the reactions and mechanisms involved is required before a reactor can be built. In general, this means the chemical reaction equilibrium thermodynamics must be known before the reactor is even conceptualized. Any chemical reaction can be written as



where A and B are the reactants, R and S the products, and a, b, r, s are defined as the stoichiometric coefficients. In general, these stoichiometric coefficients are given a value of ν_i (stoichiometric numbers). An arbitrary sign convention is given to the stoichiometric numbers to make them consistent with thermodynamics: positive signs for products, negative signs for reactants.

An example for the reaction of methane with ethylene to give butane plus hydrogen is written as



Here the stoichiometric number of methane is -2 , ethylene -1 , butane $+1$, and hydrogen $+1$. If we look at the change in the number of moles of one component, there is a direct relationship between stoichiometry and the change in the number of moles of any other component.

$$\frac{\Delta N_1}{\nu_1} = \frac{\Delta N_2}{\nu_2} = \cdots = \frac{\Delta N_i}{\nu_i}.$$

For a differential amount

$$\frac{dN_1}{\nu_1} = \frac{dN_2}{\nu_2} = \cdots = \frac{dN_i}{\nu_i} \equiv d\varepsilon,$$

where ε is the reaction coordinate or molar extent of reaction. Note that

$$\frac{dN_i}{\nu_i} \equiv d\varepsilon \quad (i = 1, 2, \dots, n)$$

This equation, with the boundary conditions,

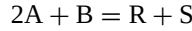
$$\begin{aligned} N &= N_{i0} & \text{for } \varepsilon = 0 \\ N &= N_i & \text{for } \varepsilon = \varepsilon \end{aligned}$$

on integration gives

$$N_i = N_{i0} + \nu_i \varepsilon \quad (i = 1, 2, \dots, n).$$

The reaction coordinate provides a relationship between the initial number of moles N_{i0} , the reaction coordinate ε , and the number of moles N_i at any point or stage in the reaction. Since the units of the stoichiometric numbers ν_i are dimensionless, the reaction coordinate has the same units as N_i (for example, mol, kg mol, or kg mol/sec).

a. *Example.* For the gas phase reaction,



7 mol of A are reacted with 4 mol of B in a batch reactor. A gas-mixture analysis after reaction showed the final mixture contained 20 mol% R. Calculate the mole fractions of the other components.

Knowns:

1. $N_{A0} = 7$.
2. $N_{B0} = 4$.
3. $Y_R = 0.20$.
4. $N_{R0} = 0$, $N_{S0} = 0$.

Let: N_T = final number of moles, N_0 = initial total number of moles, and $\nu = \sum \nu_i$.

$$\begin{aligned} N_A &= N_{A0} + \nu_A \varepsilon = 7 - 2\varepsilon \\ N_B &= N_{B0} + \nu_B \varepsilon = 4 - \varepsilon \\ N_R &= N_{R0} + \nu_R \varepsilon = 0 + \varepsilon \\ N_S &= N_{S0} + \nu_S \varepsilon = 0 + \varepsilon \\ \hline N_T &= N_0 + \nu \varepsilon = 11 - \varepsilon \end{aligned}$$

since

$$\begin{aligned} Y_A &\equiv \frac{N_A}{N_T} = \frac{7 - 2\varepsilon}{11 - \varepsilon} & Y_B &\equiv \frac{N_B}{N_T} = \frac{4 - \varepsilon}{11 - \varepsilon} \\ Y_R &= \frac{N_R}{N_T} = \frac{\varepsilon}{11 - \varepsilon} & Y_S &= \frac{N_S}{N_T} = \frac{\varepsilon}{11 - \varepsilon}, \end{aligned}$$

but

$$Y_R = \frac{N_R}{N_T} = 0.2 = \frac{\varepsilon}{11 - \varepsilon},$$

so $\varepsilon = 1.83$. Therefore, $Y_S = 0.20$ (as expected from stoichiometry) and

$$\begin{aligned} Y_S &= 0.20 \\ Y_R &= 0.20 \\ Y_A &= 0.36 \\ Y_B &= 0.24 \\ \hline \Sigma Y_i &= 1.00. \end{aligned}$$

A similar analysis can be made for many reactions occurring simultaneously. If we have r independent reactions with n species, their stoichiometric coefficients can be termed $\nu_{i,j}$, with $i = 1, 2, \dots, n$ species and $j = I,$

II, $\dots, r$ reactions. In this case,

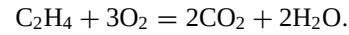
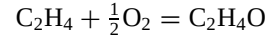
$$dN_{i,j} = \nu_{i,j} d\varepsilon_j \quad \begin{cases} i = 1, 2, \dots, n \\ j = I, II, \dots, r \end{cases}$$

and

$$dN_i = \sum_j \nu_{i,j} d\varepsilon_j \quad \begin{cases} i = 1, 2, \dots, n \\ j = I, II, \dots, r. \end{cases}$$

Integration gives $N_i = N_{i0} + \sum_j \nu_{i,j} \varepsilon_j$

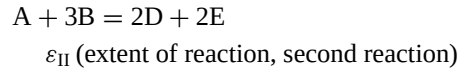
b. *Example.*



Initially, 1 mol of C_2H_4 and 3 mol of air ($\approx 21\% O_2$) react. Derive an expression relating the mole fractions of each of the components. For the reactions



and



Knowns:

1. $N_{A0} = 1$ mol.
2. $N_{B0} = (.21)(3) = 0.63$ mol.
3. $N_{C0} = N_{D0} = N_{E0} = 0$.
4. N_2 is an inert, that is, $N_{I0} = N_I = (0.79)(3) = 2.37$ mol.

In general,

$$N_i = N_{i0} + \sum_j \nu_{i,j} \varepsilon_j \quad (j = I, II)$$

$$N_A = N_{A0} + \nu_{A,I} \varepsilon_I + \nu_{A,II} \varepsilon_{II} = 1 - \varepsilon_I - \varepsilon_{II}$$

$$N_B = N_{B0} + \nu_{B,I} \varepsilon_I + \nu_{B,II} \varepsilon_{II} = 0.63 - \frac{1}{2} \varepsilon_I - 3 \varepsilon_{II}$$

$$N_C = N_{C0} + \nu_{C,I} \varepsilon_I + \nu_{C,II} \varepsilon_{II} = 0 + \varepsilon_I$$

$$N_D = N_{D0} + \nu_{D,I} \varepsilon_I + \nu_{D,II} \varepsilon_{II} = 0 + 2 \varepsilon_{II}$$

$$N_E = N_{E0} + \nu_{E,I} \varepsilon_I + \nu_{E,II} \varepsilon_{II} = 0 + 2 \varepsilon_{II}$$

$$N_I = N_{I0} + \nu_{I,I} \varepsilon_I + \nu_{I,II} \varepsilon_{II} = 2.37 + 0 \varepsilon_I + 0 \varepsilon_{II}$$

$$N_T = N_{T0} + \sum_i \sum_j \nu_{i,j} \varepsilon_j = 4 - \frac{1}{2} \varepsilon_I.$$

Therefore,

$$Y_A = \frac{1 - \varepsilon_I - \varepsilon_{II}}{4 - \frac{1}{2} \varepsilon_I}.$$

and

$$Y_B = \frac{0.63 - \frac{1}{2}\varepsilon_I - 3\varepsilon_{II}}{4 - \frac{1}{2}\varepsilon_I}.$$

Similar expressions are prepared for the remaining components.

In general, the reaction coordinate or molar extent of reaction is a bookkeeping method. Numerical values of the reaction coordinate depend on how we write the chemical reaction. When the initial moles are unknown or when preliminary calculations are done, a basis of 1 mol of feed is usually assumed. The numerical value of the reaction coordinate depends on this basis but cancels out when mole fractions are calculated.

Another commonly used method for determining the extent of reaction is conversion. Conversion is based on initial and final molar quantities of a reactant. This molar basis can be written in terms of either total moles of reactant or in terms of molar flow rate. In equation form,

$$X_A = \frac{N_{A0} - N_A}{N_{A0}},$$

where X_A is the conversion of reactant A between 0 and 1, N_{A0} the initial moles of reactant A or initial molar flow rate of A, and N_A the final number of moles or outlet molar flow rate of A.

For single reactions, fractional conversion is normally the preferred measure of the extent of reaction. However, for multiple reactions the reaction coordinate is the method of choice. The relationship that exists between conversion and the reaction coordinate is

$$X_A = -\frac{\nu_A \varepsilon}{N_{A0}}.$$

2. Rate Expressions

Before designing a chemical reactor, one must know the reaction(s) rate. Rates of reaction can be written in intrinsic form or in terms of a specific reactant of interest. An intensive measure, based on a unit volume of fluid, is normally used for homogeneous reacting systems. Thus, the general definition of reaction rate can be written as

$$r_i = \frac{1}{V^t} \left(\frac{dN_i}{dt} \right),$$

where r_i is the number of moles of component i that appear or disappear by reaction per unit volume and time in $\text{kg mol liter}^{-1} \text{sec}^{-1}$, V^t the total volume of reacting fluid in liters, N_i the number of moles of component i in kg mol, and t the time in seconds. The rates of formation of products $R, S, T, \dots$ are related to the rates of disappearance of reactants $A, B, \dots$, by the stoichiometric numbers,

$$\frac{r_A}{\nu_A} = \frac{r_B}{\nu_B} = \frac{r_R}{\nu_R} = \frac{r_S}{\nu_S} = \frac{r_T}{\nu_T} = \frac{r_i}{\nu_i}.$$

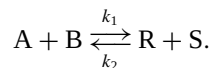
With the normal sign convention (positive for products, negative for reactants), a rate is negative for a reactant ($-r_A$) and positive for a product (r_R).

Rates of reactions are functions of the thermodynamic state of the system. For a simple system, fixing temperature and composition fixes the rest of the thermodynamic quantities or the state. Thus, the rate can be written in terms of a temperature-dependent term called the rate constant k (constant at fixed temperature) and a concentration term or terms C_i .

a. Example

$$-r_A = kC_A.$$

Rates of reaction vary with changes in temperature or concentration. All reactions are reversible (i.e., have a forward and a reverse reaction). When the rate of the forward reaction equals the rate of the reverse reaction, there is no net change in concentrations of any component, and the system is said to be at thermodynamic equilibrium. This condition represents a minimum free energy of the system and determines the limits of conversion. The overall rate of reaction equals zero at equilibrium. A relationship can be derived between the forward and reverse rate constants and the overall thermodynamic equilibrium constant. For example, consider the reaction



If the forward rate equals $k_1 C_A C_B$, and the reverse rate equals $k_2 C_R C_S$, the overall rate of disappearance of component A is $-r_A = k_1 C_A C_B - k_2 C_R C_S$. At equilibrium, $-r_A \simeq 0$,

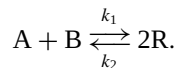
$$\frac{k_1}{k_2} = \frac{C_R C_S}{C_A C_B} \equiv K_c,$$

where K_c is defined as the thermodynamic equilibrium constant based on concentration.

Reactions that have very high values of the equilibrium constant are termed irreversible since the value of k_2 must be very small. Without much loss of accuracy, these equations can be modeled as dependent only on the forward rate. In this example, if the reaction is essentially irreversible, $-r_A = k_1 C_A C_B$.

Rate expressions must ultimately come from an analysis of experimental data. We cannot normally write a rate equation by inspection of the stoichiometric reaction equation; however, a reaction is termed elementary if the rate expression can be written by inspection based on the stoichiometric numbers.

Consider the following reversible reaction



If this reaction is elementary, the rate expression can be written as

$$-r_A = k_1 C_A C_B - k_2 C_R^2.$$

In general, an elementary reaction has the form:

$$-r_A = k_1 C_A^{|v_A|} C_B^{|v_B|} \dots - k_2 C_R^{|v_R|} C_S^{|v_S|} \dots$$

Reactions are classified by their order depending on the sum of the stoichiometric coefficients of each term.

a. Examples

$-r_A = k$	zero order
$-r_A = kC_A$	first-order irreversible
$-r_A = kC_A^2$	second-order irreversible
$-r_A = k_1 C_A - k_2 C_R$	first-order reversible
$-r_A = \frac{k_1 C_A}{1 + k_2 C_A C_R}$	complex
$-r_A = kC_A^{0.3} C_B^{0.7}$	complex

3. Use of Kinetic Data

To design a chemical reactor the rate expression must be known. Assuming the reaction is known not to be elementary, we must search for a mechanism that describes the reaction taking place or use experimental data directly. Mechanisms can be hypothesized as the sum of a series of elementary reactions with intermediates. Using methods developed by physical chemists, we can hypothesize whether the proposed mechanism fits the actual experimental evidence. If no inconsistencies are found, the hypothesized mechanism is possibly the actual mechanism. However, agreement of the mechanism with the experimental data does not necessarily mean that the proposed mechanism is correct, since many mechanisms can be hypothesized to fit such data.

An interpretation of batch or flow reactor data is used to fit an empirical rate expression. For example, in a simple batch reactor, concentration is measured as a function of time. Once the experimental data are available, two methods can be used to fit a rate expression.

The first, called the integral method of data analysis, consists of hypothesizing rate expressions and then testing the data to see if the hypothesized rate expression fits the experimental data. These types of graphing approaches are well covered in most textbooks on kinetics or reactor design.

The differential method of analysis of kinetic data deals directly with the differential rate of reaction. A mechanism

is hypothesized to obtain a rate expression and a concentration-versus-time plot is made. The equation is smoothed, and the slopes, which are the rates at each composition, are evaluated. These rates are then plotted versus concentration; and if we obtain a straight line passing through the origin, the rate equation is consistent with the data. If not, another equation is tested. Kinetic data can also be taken in flow reactors and evaluated with the above methods and the reactor design equation.

4. Temperature Dependence of the Rate Constant

On a microscopic scale, atoms and molecules travel faster and, therefore, have more collisions as the temperature of a system is increased. Since molecular collisions are the driving force for chemical reactions, more collisions give a higher rate of reaction. The kinetic theory of gases suggests an exponential increase in the number of collisions with a rise in temperature. This model fits an extremely large number of chemical reactions and is called an Arrhenius temperature dependency, or Arrhenius' law. The general form of this exponential relationship is

$$k = k_0 e^{-E/RT},$$

where k is the rate constant, k_0 the frequency factor or pre-exponential, E the activation energy, R the universal gas constant, and T the absolute temperature. For most reactions, the activation energy is positive, and the rate constant k increases with temperature. Some reactions have very little or no temperature dependence and therefore activation energy values close to zero. A few complex reactions have a net negative activation energy and actually decrease with temperature. These reactions are extremely rare.

The Arrhenius temperature dependency for a reaction can be calculated using experimental data. The procedure is to run a reaction at several different temperatures to get the rate constant k as a function of absolute temperature. From the previous equations $\ln k = \ln k_0 - E/RT$; the natural log of k is then plotted versus the reciprocal of the absolute temperature. The slope of this line is then $-E/R$, and the intercept is the $\ln k_0$.

B. Design Equations

1. General Reactor Design Equation

All chemical reactors have at least one thing in common: Chemical species are created or destroyed. In developing a general reactor design equation, we focus on what happens to the number of moles of a particular species i . Consider a region of space where chemical species flow into the

region, partially reacts, and then flows out of the region. Doing a material balance, we find rate in – rate out + rate of generation = rate of accumulation.

In equation form

$$\dot{n}_{i0} - \dot{n}_i + G_i = \frac{dN_i}{dt},$$

where $\dot{n}_{i0}$ is the molar flow rate of i in, $\dot{n}_i$ the molar flow rate of i out, G_i the rate of generation of i by chemical reaction, and dN_i/dt the rate of accumulation of i in the region. The rate of generation of i by chemical reaction is directly related to the rate of reaction by

$$G_i = \int_0^{V^t} r_i dV = \frac{dN_i}{dt}.$$

2. Ideal Batch Reactor Equation

A batch reactor has no inlet or outlet flows, so $\dot{n}_{i0} = \dot{n}_i = 0$. Perfect mixing is assumed for this ideal reactor, and the rate r_i is independent of position. This changes our generation term in the general reactor design equation to

$$\int_0^{V^t} r_i dV = r_i V^t.$$

Then, by the general design equation, our ideal batch reactor equation becomes

$$\frac{1}{V^t} \frac{dN_i}{dt} = r_i.$$

This equation does not define the rate r_i , which is an algebraic expression independent of reactor type such as $r_i = kC_i^2$.

a. Constant volume batch reactors. For the special case of constant volume or constant density (usually values for the mixture, not the reactor), we can simplify the ideal batch reactor equations. Starting with the ideal batch reactor equation

$$\frac{1}{V^t} \frac{dN_i}{dt} = r_i,$$

the volume is placed inside the differential and changed to concentration:

$$\frac{d(N_i/V^t)}{dt} = \frac{dC_i}{dt} = r_i.$$

[constant V^t , ideal batch reactor].

This equation is usually valid for liquid-phase reactions and for gas reactions where the sum of the stoichiometric numbers equals zero, but it is invalid for constant pressure gas-phase reactions with mole changes.

When the rate expression is known, this equation yields the major design variable, time, for a batch reaction of given concentration or conversion.

i. Example. $A \rightarrow B + C$ (irreversible, aqueous reaction). The rate expression can be written as

$$r_A = -kC_A.$$

Using this rate expression and the constant density ideal batch reactor equation gives

$$\frac{dC_A}{dt} = -kC_A.$$

Integrating with an initial concentration C_{A0} at $t = 0$ gives

$$\ln \frac{C_A}{C_{A0}} = -kt \quad [\text{constant volume, } V^t],$$

where t is the time for the batch reaction.

It is often convenient to work with fractional conversion of a reactant species. Let $i = A$, a reactant, then

$$X_A = \frac{N_{A0} - N_A}{N_{A0}} = \frac{N_{A0}/V^t - N_A/V^t}{N_{A0}/V^t}$$

and if V^t is constant,

$$X_A = \frac{C_{A0} - C_A}{C_{A0}} \quad [\text{constant } V^t]$$

Substituting into the ideal batch reactor equation gives

$$-C_{A0} \frac{dX_A}{dt} = r_i \quad [\text{constant } V^t]$$

ii. Example. $A \rightarrow B + C$ (elementary, constant volume reaction). The rate expression can then be written as

$$r_A = -kC_A = -kC_{A0}(1 - X_A),$$

where $C_A = C_{A0}(1 - X_A)$. Therefore,

$$-C_{A0} \frac{dX_A}{dt} = -kC_{A0}(1 - X_A).$$

Integrating with the boundary condition $X_A = 0$ at $t = 0$, gives

$$-\ln(1 - X_A) = kt \quad [\text{constant } V^t]$$

Given a rate constant k and a desired conversion, the time for the batch reaction can be calculated.

b. Variable volume batch reactors. In general, the equations developed previously assumed constant volume or constant density. For gas-phase reaction such as $A + B \rightarrow C$, the total number of moles decrease, and the volume (or density) changes.

Our ideal batch reactor equation, written in terms of any reactant A , can be changed to reflect a change in volume. For example,

$$-r_A = -\frac{dN_A}{V^t dt} = -\frac{d(C_A V^t)}{V^t dt},$$

or

$$-r_A = -\frac{1}{V^t} \left[\frac{V^t dC_A}{dt} + \frac{C_A dV^t}{dt} \right],$$

or

$$-r_A = -\left[\frac{dC_A}{dt} + \frac{C_A}{V^t} \frac{dV^t}{dt} \right].$$

From thermodynamics, assuming ideal solutions, we can derive an expression relating the volume at any conversion with the original volume,

$$V^t = V_0^t \left[1 + \left(\frac{N_{A0} \sum v_i V_i}{|v_A| V_0^t} \right) X_A \right],$$

where V_i is the molar specific volume of component i . This expression is usually simplified by defining an expansion factor in terms of any reactant; for A,

$$E_A \equiv \frac{N_{A0} \sum v_i V_i}{|v_A| V_0^t}$$

and

$$V^t = V_0^t (1 + E_A X_A).$$

This changes the ideal batch reactor equation to

$$-r_A = \frac{C_{A0}}{1 + E_A X_A} \frac{dX_A}{dt},$$

where

$$E_A \equiv \frac{C_{A0} \sum v_i V_i}{|v_A|}$$

and assumes constant temperature, pressure, and ideal solutions.

For the special case of an ideal gas mixture,

$$C_{A0} = \frac{Y_{A0} P}{RT}$$

and

$$V_i = \frac{RT}{P}$$

which leads to an easy formula to calculate the change in volume factor.

$$E_A = \frac{Y_{A0} \nu}{|v_A|},$$

where Y_{A0} is the initial mole fraction of A, ν the sum of the stoichiometric numbers, and $|v_A|$ the stoichiometric number of component A.

iii. Example. $A \rightarrow 3R$. Given that the feed is 50% A and 50% inerts, calculate E_A . By stoichiometry,

$$|v_A| = |-1| = 1$$

$$\nu = \sum v_i = 3 - 1 = 2$$

$$Y_{A0} = Y_{i0} = 0.50$$

$$E_A = \frac{Y_{A0} \nu}{|v_A|} = \frac{(0.5)(2)}{|-1|} = 1.0.$$

c. Summary of ideal batch reactor design equations.

i. General case.

$$\frac{1}{V^t} \frac{dN_i}{dt} = r_i \quad C_i \equiv \frac{N_i}{V^t}$$

$$\frac{dC_i}{dt} + \frac{C_i d(\ln V^t)}{dt} = r_i$$

$$X_A \cong \frac{N_{A0} - N_A}{N_{A0}} \quad [\text{for reactant A}]$$

$$C_{A0} \left(\frac{V_0^t}{V^t} \right) \frac{dX_A}{dt} = -r_A.$$

ii. Constant temperature and pressure ideal solution.

$$E_A \equiv \frac{C_{A0} \sum v_i V_i}{|v_A|}$$

$$\left(\frac{C_{A0}}{C_{A0} + E_A C_A} \right) \frac{dC_A}{dt} = -r_A$$

iii. Ideal gas.

$$E_A \equiv \frac{Y_{A0} \nu}{|v_A|}$$

$$\left(\frac{C_{A0}}{1 + E_A C_A} \right) \frac{dX_A}{dt} = -r_A$$

iv. Constant volume.

$$\frac{dC_i}{dt} = r_i$$

$$C_{A0} \frac{dX_A}{dt} = -r_A.$$

3. Single Ideal Flow Reactor

For batch reactors, time is the key design variable. The batch reactor design equations answer the question: How long does it take to obtain a specified conversion or concentration?

With flow reactors, volume is the key design variable. For a given feed rate, how big must the reactor be to get a specified conversion?

a. Ideal continuous-stirred tank reactor design equations. Very well-mixed unreacted material flows into a vessel, reacts, and exits the reactor along with converted product. Starting with the general reactor design equation, several assumptions are made to reduce the equation to a usable form.

$$\dot{n}_{i0} - \dot{n}_i + \int_0^{V^t} r_i dV = \frac{dN_i}{dt}$$

i. CSTR assumptions.

1. There is no accumulation in the reactor of any species i . This implies the reactor is at steady-state flow conditions.

$$\frac{dN_i}{dt} = 0$$

2. There is perfect mixing in the reactor. This implies no spatial variations of rate in the reactor, and the composition of the exit stream is the same as the composition anywhere in the reactor.

$$\int_0^{V^t} r_i dV = r_i V^t$$

These assumptions then give the ideal CSTR design equation

$$V^t = \frac{\dot{n}_i - \dot{n}_{i0}}{r_i}.$$

If this equation is written for a reactant A, the resulting equation is

$$V^t = \frac{\dot{n}_{A0} - \dot{n}_A}{-r_A}.$$

Noting that $\dot{n}_A = \dot{n}_{A0}(1 - X_A)$, we can rewrite the ideal CSTR design equation in terms of conversion of reactant A, as

$$V^t = \frac{X_A \dot{n}_{A0}}{-r_A}.$$

For the special case of constant density or constant volume of the reacting fluid, this equation is written

$$V^t = \frac{X_A \dot{n}_{A0}}{-r_A} = \frac{\dot{n}_{A0}(C_{A0} - C_A)}{C_{A0}(-r_A)}.$$

b. Ideal plug flow reactor design equation. Unreacted material flows into the reactor, a pipe or tube that has a large enough length and volume to provide sufficient residence time for the fluid to react before exiting. The assumption of ideal plug flow indicates that the composition in the reactor is independent of radial position. Unlike in a stirred-tank reactor, the composition changes as the fluid flows down the length of the reactor. The design equation for an ideal PFR is derived by a differential material balance assuming steady-state flow in the reactor. This gives

$$\dot{n}_i + \int_0^{V^t} r_i dV - (\dot{n}_i + d\dot{n}_i) = 0.$$

Upon simplification the resulting ideal plug flow reactor equation is

$$dV^t = \frac{d\dot{n}_i}{r_i}.$$

In terms of a reactant A and conversion, this equation can be written as

$$V^t = \dot{n}_{A0} \int_0^{X_A} \frac{dX_A}{-r_A}.$$

For the special case of a constant density PFR, the preceding equation can be simplified by noting that

$$X_A = \frac{C_{A0} - C_A}{C_{A0}}$$

$$dX_A = -\frac{dC_A}{C_{A0}} \quad [\text{constant volume}]$$

therefore,

$$V^t = -\frac{\dot{n}_{A0}}{C_{A0}} \int_{C_{A0}}^{C_A} \frac{dC_A}{-r_A}.$$

For the special case of a packed bed catalytic reactor with plug flow, the equation is rewritten in terms of catalyst weight,

$$W_c = \int_{\dot{n}_{i0}}^{\dot{n}_i} \frac{d\dot{n}_i}{r'_i},$$

where W_c is the weight of catalyst in kg and r'_i the rate constant based on a unit volume of catalyst in $\text{mol sec}^{-1} \text{kg}^{-1}$ catalyst.

4. Space Time

It is useful to have a measure of time for a flow reactor even though the major design variable is reactor or fluid volume. A commonly used quantity in industrial reactor design is space time. Space time is defined as the time required to process one reactor volume of feed, measured at some set of specified conditions. The normal conditions chosen are the inlet concentration of a reactant and inlet molar or volumetric flow rate.

Volumetric flow rate into the reactor is defined as

$$\dot{V}_0 \equiv \frac{\dot{n}_{A0}}{C_{A0}}.$$

Since time is obtained when total volume is divided by volumetric flow rate, a quantity τ called space time is defined

$$\tau = \frac{V^t}{\dot{V}_0} = \frac{C_{A0} V^t}{\dot{n}_{A0}}.$$

Since space time is defined for the inlet conditions, it is constant no matter what happens in the reactor. Our design

equations for a CSTR and a PFR can be modified to reflect this quantity.

$$\text{CSTR, } \tau = \frac{C_{A0}X_A}{-r_A}$$

$$\text{PFR, } \tau = C_{A0} \int_0^{X_A} \frac{dX_A}{-r_A}$$

For the special cases of constant density, these equations simplify to

$$\text{CSTR, } \tau = \frac{C_{A0} - C_A}{-r_A}$$

[constant volume or density]

$$\text{PFR, } \tau = - \int_{C_{A0}}^{C_A} \frac{dC_A}{-r_A}$$

[constant volume or density].

5. Transient Stirred-Tank Reactors

Design equations for unsteady-state operation are needed for start-up of CSTRs or for semibatch operation. These equations must have the ability to predict accurately the concentration or conversion changes before steady-state flow is obtained. Starting with the general design equation, and assuming perfect mixing, we obtain

$$\dot{n}_{i0} - \dot{n}_i + r_i V^t = \frac{dN_i}{dt}.$$

Since

$$\dot{n}_{i0} = C_{i0} \dot{V}_0$$

$$\dot{n}_i = C_i \dot{V}$$

$$N_i = C_i V^t$$

and

$$dN_i \equiv V^t dC_i + C_i dV^t$$

upon substitution the resulting equation is

$$\frac{dC_i}{dt} + \frac{C_i(\dot{V} + dV^t/dt) - C_{i0}\dot{V}_0}{V^t} - r_i = 0.$$

C. Design Considerations

1. Batch Versus Flow Reactors

Commercial-scale batch reactors are generally used for small-lot or specialty items. This includes chemicals such as paints, dyes, and pharmaceuticals. Batch reactors are very simple and flexible. Vessels used to make one compound can be washed and reused to make other products. The ease of cleaning and maintaining batch reactors along with low capital investment and low instrumental costs

make the batch reactor particularly attractive in industrial applications.

The batch reactor also has disadvantages. These include high labor cost, manual control, poor heat transfer conditions, and mixing problems. Poor heat transfer results from relatively low area-to-volume ratios. This can be avoided with the use of internal coils or external recycle heat exchangers. Batch reactors are generally not suitable for highly endothermic or highly exothermic reactions. These heat effects can be partially avoided by running in a semibatch operation.

Good mixing is required for approaching theoretical conversion. Depending on impeller design, a power of 0.5–1.0 kW/m³ produces 90% of the calculated theoretical conversion. Care must be taken to design batch reactors with a height-to-diameter ratio close to one. For larger ratios, pump circulation or baffling is required. For high-pressure reactions, sealing problems may be encountered on the agitator shaft.

Perhaps the biggest disadvantage of a batch reactor is the difficulty encountered for isolation of intermediates. For series reactions such as $A \rightarrow B \rightarrow C$, where B is the desired product, it is difficult to stop the reaction (quench) without overshooting.

Continuous tubular flow reactors are most commonly used for large quantity items such as chemicals manufactured in the petroleum industry. There are many advantages of continuous tubular flow reactors. Labor costs are very low, and automatic control is easy to implement. Liquid- or gas-phase homogeneous reactions are routine for all temperature and pressure ranges. Heterogeneous reactions, such as solid-catalyzed reactions, are easily run in packed beds or packed tube reactors. Intermediates are easy to isolate for any desired conversion, since the reactor length can be adjusted. Heat transfer is relatively good with large area-to-volume ratios and can be made as large as required by using smaller tubes. For large heat effects, the reactor can be designed as a counter-current heat exchanger or as a single jacketed reactor. For highly endothermic reactions, the reactor tubes can be placed in a furnace and heated radiantly or with hot combustion gases.

Tubular flow reactors are usually inflexible. Normally they are designed and dedicated to a single process. They are typically hard to clean and maintain, have high capital costs, and depending on materials and geometry, are rarely stock items.

To achieve desired conversions predicted by ideal design equations, plug flow is required. This implies turbulent flow and higher energy costs if packing is used. Mass transfer can also be a problem. Axial diffusion or dispersion tends to decrease residence time in the reactor. High values of the length-to-diameter ratios ($L/D > 100$) tend to minimize this problem and also help heat transfer.

Continuous-stirred tank reactors lie somewhere between tubular and batch reactors. Mixing and heat transfer problems are similar to those of batch reactors. However, many of the stirred-tank reactors have benefits of the tubular flow reactors. These include isolation of intermediates, automatic control, and low labor costs.

2. Heat Effects

Most reactors used in industrial operations run isothermally. For adiabatic operation, principles of thermodynamics are combined with reactor design equations to predict conversion with changing temperature. Rates of reaction normally increase with temperature, but chemical equilibrium must be checked to determine ultimate levels of conversion. The search for an optimum isothermal temperature is common for series or parallel reactions, since the rate constants change differently for each reaction. Special operating conditions must be considered for any highly endothermic or exothermic reaction.

3. Design for Multiple Reactors

Common design problems encountered in industrial operations include size comparisons for single reactors, multiple reactor systems, and recycle reactors.

a. Size comparisons of single isothermal flow reactors. The rate of reaction of a CSTR is always fixed by the outlet concentrations. Since the rate is constant (first- or second-order, etc.), a large volume is required to provide enough time for high conversion. In general, a plug flow reactor is much more efficient and requires less volume than a stirred-tank reactor to achieve the same level of conversion. In a plug flow reactor, the rate changes down the length of the reactor due to changes in reactant concentrations. High initial rates prevail in the front of the reactor with decreasing rates near the end. The overall integration of these rates is much higher than the fixed rate in a CSTR of equal volume. For complex kinetics such as autocatalytic reactions, where the concentrations of both reactants and products increase the forward rate of reaction, stirred-tank reactors are preferred and require less volume. Under most common kinetics, a series of three or four stirred-tank reactors of equal volume in series approaches the performance of a plug flow reactor.

b. Reactors in series and parallel.

i. Plug flow reactors. Plug flow reactors are unique in the sense that operation in parallel or series give the same conversion if the space time is held constant. This implies, for example, that if a 20-m reactor of fixed diameter is required to achieve a given conversion, the same

conversion and capacity can be achieved by running ten 2-m reactors in series or ten 2-m reactors in parallel. The split of the feed in the parallel case must be one tenth of the total to keep the same space time. In industrial applications the geometry chosen is a function of cost of construction, ease of operation, and pressure drop. Parallel operation is normally preferred to keep the pressure drop at a minimum.

ii. Stirred-tank reactors in series and parallel.

Stirred-tank reactors behave somewhat differently from plug flow reactors. Operation of CSTRs in parallel, assuming equal space time per reactor, gives the same conversion as a single reactor but increases the throughput or capacity proportional to the number of reactors.

This is not the case for multiple CSTRs in series. CSTRs operated in series approach plug flow and therefore give much higher levels of throughput for the same conversion. When we have two reactors of unequal size in series, highest conversion is achieved by keeping the intermediate concentration as high as possible. This implies putting the small CSTR before the large CSTR.

c. Plug flow and stirred-tank reactors in series.

When two reactors, a plug flow and a stirred tank are operated in series, which one should go first for maximum conversion? To solve this problem the intermediate conversion is calculated, the outlet conversions are determined, and the best arrangement chosen. Keeping the intermediate conversion as high as possible results in the maximum conversion. Concentration levels in the feed do not affect the results of this analysis as long as we have equal molar feed.

4. Recycle Reactors

In a recycle reactor, part of the exit stream is recycled back to the inlet of the reactor. For a stirred-tank reactor, recycle has no effect on conversion, since we are essentially just mixing a mixed reactor. For a plug flow reactor, the effect of recycle is to approach the performance of a CSTR. This is advantageous for certain applications such as autocatalytic reactions and multiple reaction situations where we have a PFR but really require a CSTR.

IV. SPECIAL REACTOR CONFIGURATIONS

Additional reactors exist that are either completely or partially based on the five primary reactor types discussed in Section II. They receive special attention due to specific applications and/or unique mass transfer characteristics.

A. Autoclave

1. Description

The autoclave reactor is a small cylindrical reactor, built to withstand high pressures, used to evaluate the kinetics of high-temperature, high-pressure reactions and the production of small quantities of specialty chemicals. The reactor is typically packed with a supported catalyst, and reactant is added by injection. Pressure in the system is elevated by increasing the temperature of the autoclave. Additional pressure, if needed, can be obtained with the injection of additional gaseous reactant or an inert.

2. Classification

The autoclave is usually a heterogeneous batch reactor mainly used for high-pressure kinetic studies. The autoclave is typically a solid catalyzed gas–liquid reaction system.

3. Applications

This reactor allows easy data collection for high-temperature, high-pressure reaction systems that have difficult flow properties. This includes reactants that are solid at room temperature or mixtures of solids and liquids. Typical reactions performed in autoclaves are coal liquefaction, petroleum residuals and coal liquids upgrading, and high molecular weight hydrogenation experiments.

B. Blast Furnace

1. Description

The blast furnace, a vertical shaft kiln, is the oldest industrial furnace. Reactant enters in the top of the shaft and falls down through a preheating section, a calcinating section, past oil, gas, or pulverized coal burners, through a cooling section, with the product ash falling through a discharge gate.

2. Classification

The blast furnace operates continuously although the individual particles see a batch mode of reaction. The actual reaction conditions must be based on the batch reactor sequence for the particles since complete conversion is desired. This requires control of the mass throughput in the furnace, but primarily it requires accurate temperature control. Control of the solids is maintained at the bottom discharge port. Gas flow rate is controlled by blowers or by a stack discharge fan.

3. Applications

Blast furnaces are used for the production of iron from ore and phosphorus from phosphate rock.

C. Bubble Column

1. Description

The bubble column is a tower containing primarily liquid (>90%) that has a gas or a lighter liquid sparged into the bottom, allowing bubbles to rise through the column. The column may contain staging, which enhances the mass transfer characteristics of the reactor. In countercurrent operation the reactor is particularly attractive for slightly soluble gases and liquid–liquid systems. With cocurrent flow and a highly baffled column, the reactor has mass transfer characteristics similar to those of a static mixer. The reactor may sometimes contain a solid suspended in the liquid phase.

2. Classification

The bubble column is a typical gas–liquid heterogeneous reactor with the design also applicable to liquid–liquid systems. The bubbles rise through the liquid in plug flow. The liquid is well mixed by the bubbling gas and seldom follows plug flow assumptions.

3. Applications

The bubble column can withstand high gas velocities and still maintain high mass transfer coefficients. This column is particularly attractive for reactions that do not require large amounts of gas absorption or require well-mixed liquids.

There are numerous applications for bubble columns, for example, gas–liquid columns include the absorption of isobutylene in sulfuric acid, and liquid–liquid columns are used for nitration of aromatic hydrocarbons.

D. Chemostat–Turbidostat

1. Description

The chemostat is a biological CSTR where the substrate concentration in the tank is maintained constant. The turbidostat is similar to the chemostat except that the cell mass in the reactor is kept constant. The primary distinction between the two reactors is the control mechanism used to maintain continuous operation. A unique feature of a biological CSTR is the washout point. When the flow rate is increased so that the microbes can no longer reproduce fast enough to maintain a population, the microbes wash out of the tank, and the reaction ceases. This washout point represents the limits of maximum flow rate for operation.

2. Classification

The chemostat is a biological heterogeneous CSTR. The microbes are considered a solid phase, and for aerobic

fermentations, oxygen or air is bubbled through the tank to allow oxygen mass transfer into the media, resulting in a three-phase reactor.

3. Applications

Continuous fermentation processes are primarily used in the research and development stage. However, more chemostat operations are being used at the production level as the understanding of this reactor increases. Examples include ethanol fermentation for the production of fuel grade ethanol and single-cell protein production from methanol substrates.

E. Digestor

1. Description

The digester is a biological reactor used mainly for the treatment of municipal and industrial wastes. Wastes are fed continuously to the digester, where some solids settle to the bottom of the tank, and other solids are matted and lifted to the surface by the gases produced during the fermentation. In an aerobic digester the mat is broken and mixed by gas circulation. The solid sludge in the bottom of the tank is raked down a conical bottom and pumped from the tank. A fraction of the sludge is recycled back to the digester to maintain a steady microbial population.

2. Classification

The digester is classified as a continuous biological heterogeneous reactor. Liquid flow through the digester roughly follows the CSTR assumptions. Digestion of the solids is a complex mechanism that requires empirical design equations to describe.

3. Applications

The digester is mainly restricted to the treatment of municipal and industrial wastes. Substantial research has been done on using anaerobic digestion of biomass for the production of methane gas. These systems are limited to small-scale applications where alternative energy sources are inadequate. Some current anaerobic digesters use the methane produced as a by-product to supply heat for operation of the digester.

F. Extruder

1. Description

For reactions that require high temperature and pressure for short periods of time, the extruder is ideal. The reactant is fed to a screw type device that narrows toward the exit.

Friction in the extruder produces high temperatures and pressures, and the product is forced out dies at the end of the extrusion tube. This type of extruder is referred to as a dry extruder. If steam is injected along the extrusion tube, the reactor is referred to as a wet extruder.

2. Classification

The extruder is essentially a plug flow reactor. Although the material is being well mixed, this mixing is primarily in the radial and circumferential directions rather than axially. Due to the extreme conditions in the extruder, solids can liquefy, resulting in heterogeneous operation.

3. Applications

The extruder is used extensively in the food processing industry. Grains and starches can be hydrolyzed easily.

G. Falling Film

1. Description

Falling-film reactors have a liquid reactant flowing down the walls of a tube with a gaseous reactant flowing up or down (usually countercurrent). This reactor is particularly advantageous when the heat of reaction is high. The reaction surface area is minimal, and the total reaction heat generated can be controlled.

2. Classification

This reactor may follow the plug flow assumptions, or it may be equilibrium limited depending on the operating conditions.

3. Applications

An example of a reaction performed in a falling-film reactor is the sulfonation of dodecyl benzene.

H. Fermentor

1. Description

The term fermentation is used to describe the biological transformation of chemicals. In its most generic application, a fermentor may be batch, continuous-stirred tank (chemostat), or continuous plug flow (immobilized cell). Most industrial fermentors are batch. Several configurations exist for these batch reactors to facilitate aeration. These include sparged tanks, horizontal fermentors, and biological towers.

2. Classification

The most traditional application of the fermentor is in batch mode. In anaerobic fermentations the reactor looks like a normal batch reactor, since gas–liquid contact is not an important design consideration. Depending on the reaction, the microbes may or may not be considered as a separate phase. For aerobic fermentations, oxygen is bubbled through the media, and mass transfer between phases becomes one of the major design factors.

3. Applications

Since the characteristics of microbes lead to the batch production of many products, examples of fermentors are numerous. They include beer vats, wine casks, and cheese crates as anaerobic food production equipment. The most significant aerobic reactor is the penicillin fermentor.

I. Gasifiers

1. Description

A gasifier is used to produce synthesis gas from carbonaceous material. The solid is packed in a column, and gas is passed up through the bed, producing a mixture of combustible products, primarily methane, hydrogen and carbon monoxide, with a low to medium BTU content.

2. Classification

A gasifier is a continuous gas process in conjunction with either a batch of solids or continuous solids feed and product removal. The gas phase passing up through the bed obeys plug flow behavior. In continuous solids handling, the bed is fed from the top and emptied from the bottom. These solids also obey plug flow assumptions with flow countercurrent to the gas phase.

3. Applications

Coal gasifiers are used for the production of synthesis gas; however, any carbon source could be used. Biomass is receiving attention as a carbon source.

J. Immobilized Cell

1. Description

The washout problems associated with continuous fermentation are eliminated by attaching the microbes or enzymes to a solid support, preventing them from leaving the reactor. The attachment procedures vary, and as a

result, the flow scheme in the reactor may differ depending on the choice of attachment. Encapsulation allows shear at the surface of the support so that fluidization techniques can be used. Attachment onto a surface usually limits the flow conditions to a packed-bed configuration.

2. Classification

An immobilized cell reactor is classified as a continuous biological system that may follow either plug flow theory or fluidized-bed theory depending on the mode of operation.

3. Applications

The use of immobilized cell systems is applicable to all fermentation schemes and is being researched extensively for the production of alcohols, chemicals, and biological products.

K. Jet Tube

1. Description

For rapid exothermic reactions that require continuous stirred-tank operating conditions for good reaction control, a jet tube reactor can be used. This reactor gives thorough mixing despite the extremely short residence times involved in these reactions. One reactant is injected into the other through a jet, orifice, or venturi. This gives high turbulence to insure a well-mixed condition. Large-scale testing is needed to select the reactor conditions accurately, since minor errors in kinetic constants are magnified due to the high activation energies of the reactions. Jets can handle both gas and liquid feed and are usually built in multiple jet configurations.

2. Classification

Since reaction does not occur until one reactant is jetted into the other, the actual jet does not become involved in the kinetics, it is strictly a method for contacting reactants quickly. The actual reactor performance is based on CSTR assumptions.

3. Applications

Oil burners are jet tube reactors. Jet washers are used for fast reactions such as acid–base reactions. An example is the absorption of hydrochloric acid in sodium hydroxide–sodium sulfite solutions.

L. Lagoon

1. Description

Lagoons are used for the deposition and degradation of industrial and human wastes. The waste, in water, is pumped into a holding lagoon. Water in the lagoon usually evaporates but may be pumped out under some conditions. The advantage of a biological lagoon is long holding times for the degradation of compounds that have extremely slow reaction rates.

There are three modes of operation for lagoons. They may be either anaerobic, aerobic, or facultative (which is a combination of aerobic and anaerobic). Aerobic lagoons require the additional cost of aerators and compressors for continuous bubbling of air, oxygen, or ozone into the lagoon.

2. Classification

The biological lagoon is difficult to categorize since a reaction and a separation process are occurring simultaneously. Water flow through the system should ideally be at steady state; however, variable input, climatic conditions, and rain all affect the water in the system as a function of time. Chemical concentrations are similar to semibatch operation but may be at a relatively steady state.

3. Applications

Lagoons are a simple, low-cost reaction system for wastewater treatment. Anaerobic lagoons are capable of handling high-concentration wastes but then require an aeration lagoon to treat the water effluent. Effluent from aerobic lagoons with low-concentration feed usually requires no additional treatment to meet water quality standards.

M. Loop Reactor

1. Description

For reactions where high-pressure requirements do not allow large diameter tanks for homogeneous reaction kinetics, a loop reactor can be used. The loop is a recycle reactor made of small diameter tubes. Feed can be supplied continuously at one location in the loop and product withdrawal at another.

2. Classification

Despite its complex construction, the loop is essentially a stirred-tank reactor. By recirculating fast enough the system can be considered well mixed. For this to be the case,

the rate of recycle must be much greater than the rate of product withdrawal.

3. Applications

An example for the loop reactor is the oxidation of normal butane.

N. Packed Bed

1. Description

The packed bed reactor is used to contact fluids with solids. It is one of the most widely used industrial reactors and may or may not be catalytic. The bed is usually a column with the actual dimensions influenced by temperature and pressure drop in addition to the reaction kinetics. Heat limitations may require a small diameter tube, in which case total through-put requirements are maintained by the use of multiple tubes. This reduces the effect of hot spots in the reactor. For catalytic packed beds, regeneration is a problem for continuous operation. If a catalyst with a short life is required, then shifting between two columns may be necessary to maintain continuous operation.

2. Classification

A packed bed reactor is a continuous heterogeneous reactor. The gas or liquid phase obeys plug flow theory. The solids are considered batch, with even long-life catalyst beds losing activity over time.

3. Applications

Noncatalytic packed bed reactors have been discussed separately in other sections of this article. They include blast furnaces, convertors, roasting furnaces, rotary kilns, and gasifiers.

O. Recycle

1. Description

A recycle reactor is a mode of operation for the plug flow reactor in reaction engineering terms. Recycle may also be used in other configurations involving a separation step. In plug flow some percentage of the effluent from the reactor is mixed back into the feed stream. The reason for this is to control certain desirable reaction kinetics. The more recycle in a plug flow reactor, the closer the operation is to a stirred-tank reactor. Therefore, with recycle it is possible to operate at any condition between the values predicted by either CSTR or PFR. There is no advantage in operating

a CSTR with recycle unless a separation or other process is being performed on the recycle stream, since the CSTR is already well mixed.

2. Classification

The recycle reactor is used to reach an operating condition between the theoretical boundaries predicted by the continuous stirred tank reactor and the plug flow reactor.

3. Applications

The recycle reactor is used to control the reaction kinetics of multiple reaction systems. By controlling the concentration present in the reactor, one can shift selectivity toward a more desired product for nonlinear reaction kinetics.

P. Roasting Furnace

1. Description

Roasting furnaces are in a class of reactors used by the metallurgical industry in a preparatory step for the conversion of ores to metals. There are three widely used roasted furnaces: multiple hearth, fluidized bed, and flash roasters. In the multiple hearth configuration hot gases pass over beds of ore concentrate. The flash roaster injects pulverized ore with air into a hot combustion chamber. The fluidized bed roaster operates as described in a separate heading.

2. Classification

All of these roasting furnace reactors operate continuously. They are noncatalytic gas–solid heterogeneous reactors. The multiple hearth has characteristics similar to plug flow operation. The flash roaster approaches CSTR, and the third option is a fluidized bed configuration.

3. Applications

Roasting furnaces are used to react sulfides to produce metal oxides, which can be converted to metals in the next process step. The sulfides are used as a reducing agent in nonferrous metallurgy for the recovery of metals. The process has been used for metals such as copper, lead, zinc, nickel, magnesium, tin, antimony, and titanium.

Q. Rotary Kilns

1. Description

The rotary kiln is a long tube that is positioned at an angle near horizontal and is rotated. The angle and the rotation allow solid reactants to work their way down the tube.

Speed and angle dictate the retention time in the kiln. Gas is passed through the tube countercurrent to the solid reactant. The kiln is operated at high temperatures with three or four heating zones depending on whether a wet or dry feed is used. These zones are drying, heating, reaction, and soaking. Bed depth is controlled at any location in the tube with the use of a ring dam.

2. Classification

The rotary kiln is a continuous countercurrent heterogeneous reactor. Solids traveling down the kiln are in plug flow, as are the gases passing upward.

3. Applications

The most common reactor of this type is the lime kiln. This is a noncatalytic reaction where gas reacts with calcium carbonate moving down the kiln. Other reactions performed in the rotary kiln include calcination, oxidation, and chloridization.

Use of rotary kilns for hazardous waste incineration is becoming more common for disposal of chlorinated hydrocarbons such as polychlorinated biphenyls (PCBs). Flow in these kilns is cocurrent. Major advantages include high temperature, long residence time, and flexibility to process gas, liquid, solid, or drummed wastes.

R. Slurry Tank

1. Description

The slurry tank is a three-phase reactor where gas is bubbled up through a liquid–solid mixture. The slurry tank has the advantage of uniform temperature throughout the mixture. This temperature control is extremely important for highly exothermic reactions. Another advantage of the slurry tank is the low intraparticle diffusion resistance for this contacting pattern. As a disadvantage, low mass transfer rates occur in liquids when compared with gases, requiring that small solid particles be used. These particles can clog screens in the effluent stream used to keep solids in the tank, thus making catalyst retention difficult.

2. Classification

The slurry tank, when well mixed, can be considered a continuous-stirred tank reactor for both the gas phase and the liquid phase. When the solid is retained in the reaction vessel, it behaves in a batch mode; however, catalyst can be removed and regenerated easily in a slurry tank, so activity can be maintained.

3. Applications

A major application of the slurry tank is the polymerization of ethylene. Gaseous ethylene is bubbled through a slurry of solvent and polymer.

S. Spray Towers

1. Description

A spray tower is a continuous gas–liquid reactor. Gases pass upward through a column and contact liquid reactant sprayed into the column. The spray tower represents the opposite extreme from a bubble tower. The spray tower has greater than 90% of the volume as gas. This allows for much reduced liquid-handling rates for highly soluble reactants.

2. Classification

The spray tower is a heterogeneous gas–liquid reactor. The gas passing up the column obeys plug flow conditions, and the liquid sprayed into the column behaves either as plug flow or as batch for individual droplets falling down the tower.

3. Applications

Spray towers can be used to absorb gaseous reactants. The most widely used spray tower is for flue gas desulfurization. SO_2 in a combustion gas is passed upward through an alkaline solution that usually contains calcium oxide. The SO_2 is absorbed into the liquid, which then reacts to calcium sulfite and continues on to calcium sulfate.

T. Trickle Bed

1. Description

A trickle bed is a continuous three-phase reactor. Three phases are normally needed when one reactant is too volatile to force into the liquid phase or too nonvolatile to vaporize. Operation of a trickle bed is limited to cocurrent downflow to allow the vapor to force the liquid down the column. This contacting pattern gives good interaction between the gaseous and liquid reactants on the catalyst surface.

2. Classification

The trickle bed reactor allows for plug flow reactor assumptions even at extremely low liquid-flow rates. The trickle bed is classified as a continuous heterogeneous catalytic reactor.

3. Applications

This reactor also allows for easy laboratory scale operation for determining rate data, since the flow rate is low. Experimental-scale trickle beds can be on the order of 0.5 in. in diameter. Trickle bed reactors are used for the hydrodesulfurization of liquid petroleum fractions.

SEE ALSO THE FOLLOWING ARTICLES

ATOMIC AND MOLECULAR COLLISIONS • BATCH PROCESSING • CHEMICAL THERMODYNAMICS • FLUID MIXING • HEAT TRANSFER

BIBLIOGRAPHY

- Blanch, H. W., and Clark, D. S. (1997). "Biochemical Engineering," Marcel Dekker, New York.
- Duncan, T. M., and Reimer, J. A. (1998). "Chemical Engineering Design and Analysis: An Introduction," Cambridge University, U.K.
- Fogler, S. H. (1998). "Elements of Chemical Reaction Engineering," Prentice-Hall PTR, Englewood Cliffs, NJ.
- Hortacsu, Ö., Rippin, D., and Suno, W. T. (1996). "Batch Processing Systems Engineering: Fundamentals and Applications for Chemical Engineering," Springer-Verlag, New York.
- Levenspiel, O. (1998). "Chemical Reaction Engineering," 3rd ed., Wiley, New York.
- Peacock, D. G., and Richardson, J. F. (1999). "Chemical Engineering, Volume 3," Chemical and Biochemical Reactors & Process Control, Butterworth-Heinemann, Woburn, MA.
- Perry, R., and Green, D. (1999). "Perry's Chemical Engineering Handbook on CD-ROM," McGraw-Hill Professional, New York.
- Perry, R., Green, D., and Dean, J. (1999). "Perry's Deluxe Suite of Chemical and Chemical Engineering Data," McGraw-Hill Professional, New York.
- Rohr, Ph. R. (1996). "High Pressure Chemical Engineering," Elsevier, New York.
- Smith, J. M., and Van Ness, H. (1996). "Intro to Chemical Engineering Thermodynamics," 5th ed., McGraw-Hill Higher Education, New York.
- Tassios, D. P. (1993). "Applied Chemical Engineering Thermodynamics," Springer-Verlag, New York.
- Tominaga, H. (ed.). (1998). "Chemical Reaction and Reactor Design," Wiley, New York.



Solvent Extraction

Teh C. Lo

T. C. Lo & Associates

M. H. I. Baird

McMaster University

- I. General Principles
- II. Industrial Extraction Equipment
- III. Industrial Extraction Processes
- IV. Recent Advances

GLOSSARY

Axial mixing Eddy diffusion in the direction of the axis of the extractor and a radial diffusion or spreading, resulting from nonuniform velocity.

Countercurrent Method of extraction such that the feed solution and the solvent flow in opposite directions.

Extract Solution containing the desired product, resulting from an extraction process.

Extractant Substance added to the solvent in order to enhance the extraction process.

Feed Initial solution subjected to an extraction process and containing desired product.

Flooding Hydrodynamic instability occurring in continuous countercurrent extraction due to excessive flow rates supplied to the process.

Fractional extraction Countercurrent extraction using two solvents to separate a mixture of two or more solutes.

Membrane A thin film of liquid held between two liquid phases which are both immiscible with the membrane liquid.

Raffinate That part of the feed solution remaining after extraction of the desired product.

Solvent Liquid brought into contact with the feed to extract the desired product in extraction process.

Stage Idealized extraction process in which the feed and solvent are brought to equilibrium and then separated as raffinate and extract.

Supercritical extraction Extraction at high pressures by means of a supercritical fluid which becomes a gas when pressure is reduced below its critical pressure.

SOLVENT extraction (liquid–liquid extraction) is the separation and/or concentration of the components of a solution by distribution between two immiscible liquid phases. A particularly valuable feature is its power to separate mixtures into components according to their chemical type. Solvent extraction is widely used in the chemical industry. Its applications range from hydrometallurgy, e.g., reprocessing of spent nuclear fuel, to fertilizer manufacture and from petrochemicals to pharmaceutical products. Important factors in industrial extraction are the selection of an appropriate solvent and the design of equipment most suited to the process requirements.

I. GENERAL PRINCIPLES

A. Equilibrium in Extraction Systems

Extraction systems of practical interest will contain, in the simplest case, three components. There are two immiscible or very slightly miscible solvents (here denoted A and B) and a solute (C) that is to be extracted. The initial feed solution consists of C dissolved in A, while the extracting solvent is taken to be B. In a single extraction stage as shown schematically in Fig. 1, the feed and the solvent are first brought to equilibrium by prolonged contact usually assisted by mechanical agitation, e.g., shaking or stirring. The phases are then permitted to separate by virtue of their different densities, providing an extract (mainly B plus most of C) and a raffinate (residual amounts of C dissolved in A). In the event that the solvents are partially miscible, small amounts of A and B will be present in the extract and the raffinate, respectively.

The distribution ratio m is defined as the ratio of mass fractions of C in each phase at equilibrium:

$$m = x_{CB}/x_{CA}. \quad (1)$$

Alternatively, it can be expressed as a ratio of concentrations or mole fractions. The solvent B should be selected so that m is as large as possible yet consistent with other factors such as cost and safety.

In general, m is not independent of composition, and often the solvents A and B show partial miscibility. Ternary equilibrium data of this type may be shown on a triangular composition diagram (Fig. 2). It will be seen that in parts

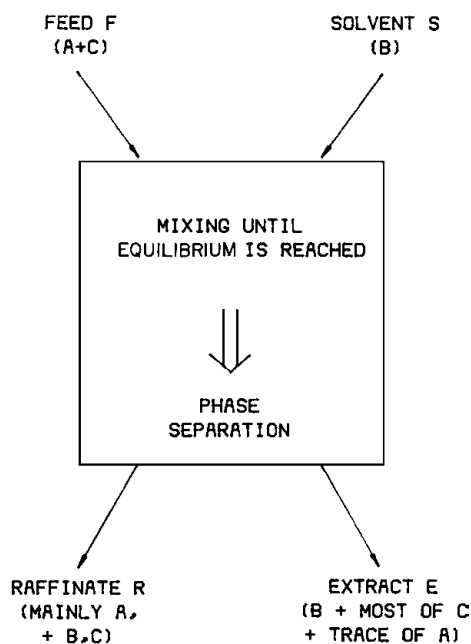


FIGURE 1 Single extraction stage.

of this diagram, no phase separation can occur. Solvent extraction is only possible if the mixed composition of the feed and the solvent lies on a point within the two-phase region.

In the graphical example shown in Fig. 2, interphase equilibria are shown by dashed tie-lines connecting the raffinate and extract compositions. As the mass fraction of C is increased, the tie-lines become shorter until the limit of miscibility is reached at the point P on Fig. 2.

The inverse lever rule indicates that when feed and solvent are mixed, their average composition lies on a point M (Fig. 2) such that M lies on a straight line between the points F (feed composition) and S (solvent composition) and that

$$\frac{\text{Distance FM}}{\text{Distance MS}} = \frac{\text{Mass of solvent added}}{\text{Mass of feed added}}. \quad (2)$$

In the example of Fig. 2, the solvent/feed ratio is 1.5.

Since the point M lies in the two-phase region of the triangular diagram, the term “mixture” applies only on a scale larger than the size of the droplets formed. The droplet dispersion formed by agitation has sufficient interfacial area (see Section I.C) for equilibrium to be reached quickly, so that point M represents the mean of the extract composition (point E) and the raffinate composition (point R) which are connected by the appropriate tie-line. A further application of the inverse lever rule permits calculation of the relative amounts of extract and raffinate. In this example, the material balance based on 1 kg of feed is summarized as follows:

Component	Feed in	Solvent in	Raffinate out	Extract out
A	0.80	—	0.740	0.060
B	—	1.50	0.005	1.495
C	0.20	—	0.015	0.185
Total	1.0 kg	1.50 kg	0.76 kg	1.74 kg

An extraction process should take as much of the solute as possible into the extract phase. This objective is expressed as the extraction factor, the ratio of the mass of C in the extract to that in the raffinate, for the single stage (12.3 in the above example). The extraction factor is increased by using a high ratio of solvent to feed and by choosing a system with a high distribution ratio. For the special case of a very dilute system with immiscible A and B and constant distribution ratio, it can be shown that the extraction factor is given by

$$\varepsilon = m(M_B/M_A), \quad (3)$$

where M_A and M_B are the masses of the components A and B fed to the equilibrium stage in the feed and solvent phases.

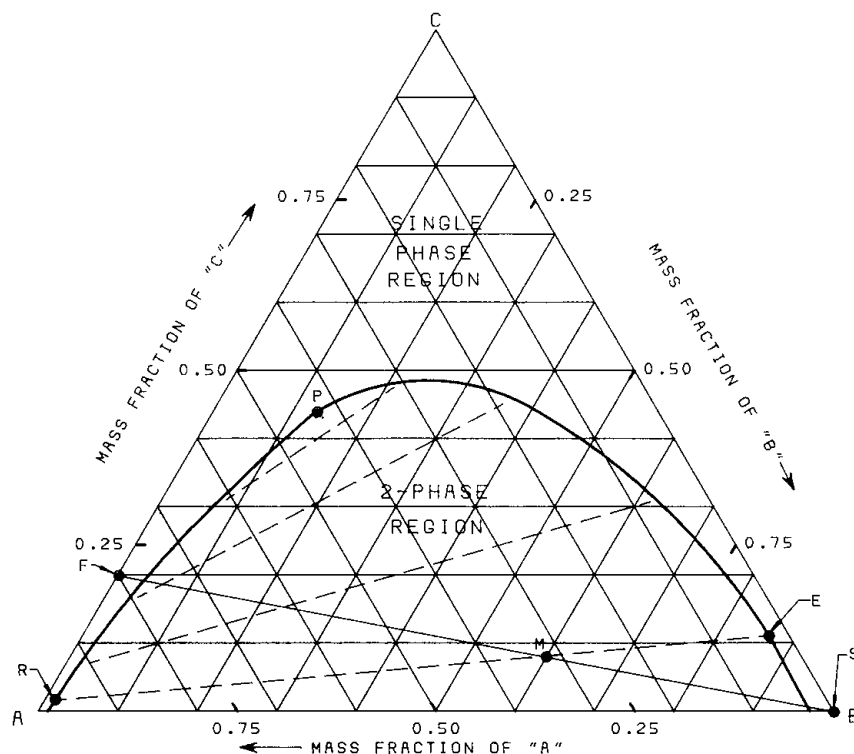


FIGURE 2 Triangular equilibrium diagram showing extraction of feed F by solvent S to give extract E and raffinate R. Dashed lines are tie-lines.

An extraction process may be required to separate two solutes C and D. In this case, the selectivity β_{CD} should be as high as possible, where

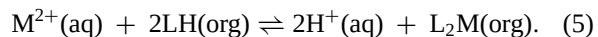
$$\beta_{CD} = \frac{m_C}{m_D} = \frac{x_{CB} x_{DA}}{x_{CA} x_{DB}}. \quad (4)$$

For dilute solutions the selectivity may be assumed to be independent of composition.

Equilibrium data have been obtained experimentally for many systems, and literature sources should be carefully checked before a decision is reached to perform an experimental measurement. The reader should also refer to published data banks from which parameters for the NRTL or UNIQUAC equations can be obtained. Recently, significant progress has been made in estimating equilibrium data by computer simulations of the molecular dynamics.

Solvent extraction may be accompanied by a chemical reaction. The selectivity and the extraction factor can be greatly improved by carrying out the extraction with a solution of an extractant that chemically converts the solute to a form that is preferentially soluble in the extracting solvent. An additional advantage of this procedure is that the reverse extraction of solute (stripping) can often be carried out by changing the equilibrium constant of the reaction, e.g., by changing the pH or temperature.

A well-known example of this type of extraction is the purification of hydrometallurgical leach solutions containing copper, nickel, etc. A metal cation M^{2+} present in the aqueous phase reacts selectively at the interface with a complexing agent dissolved in the organic solvent (e.g., kerosene). The complexing agent may have the molecular form LH with a free hydrogen atom, e.g., a carboxylic acid, an organophosphoric acid, or a hydroxyoxime. The reaction equilibrium takes the overall form



The extractant is typically present in 5–20% (by mass) concentration and is selected to give almost complete extraction at pH 4 or above. The metal species M may subsequently be stripped to the aqueous phase in purified and enriched form using a dilute mineral acid which drives the equilibrium in Eq. (5) to the left. While Eq. (5) refers to cationic species, anionic species can be extracted with solutions of amines. Stripping is carried out with strong aqueous alkali.

Another category of reactive extraction involves irreversible reactions, as in the saponification of esters (soap manufacture, etc.). Recently it has been found that equilibria can be affected by the formation of micelles, which are small clusters of molecules with diameter in the order

of a few nanometers. The micelles are stabilized by means of surfactants and extraction can be promoted in this way, provided the conditions are carefully controlled.

B. Rates of Mass Transfer

The rate at which equilibrium is reached in a given system is usually expressed as the mass transfer rate of the solute, in units of mass per unit time. The mass transfer rate in the absence of a chemical reaction is proportional to the product of the interfacial area and the solute concentration driving force (departure from equilibrium). The proportionality constant is known as a mass transfer coefficient, which is nearly always a function of the molecular diffusivity of the solute. In this section the concepts of mass transfer coefficient and concentration driving force are briefly reviewed.

For mass transfer in a simple ternary system without chemical reaction, the solute concentration profiles near the interface are as shown in Fig. 3. The concentration in the bulk of each phase is uniform because of convective mixing effects, but very near the interface the rate of mass transfer depends increasingly on molecular diffusion.

The combined effects of diffusion and convective mixing are included in the mass transfer coefficients k_A and k_B , which relate flux to concentration difference in the interfacial region of each phase,

$$N = k_A(c_A - c_{Ai}) = k_B(c_{Bi} - c_B). \quad (6)$$

The thicknesses of the interfacial regions across which the concentrations vary are typically in the order of

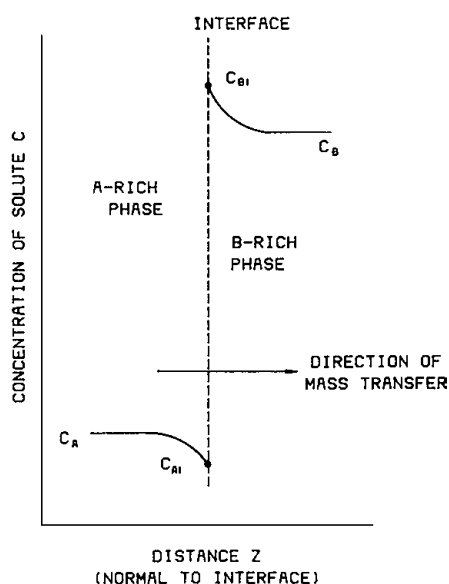


FIGURE 3 Solute concentration profiles near interface in extraction process.

100 μm . It is therefore very difficult to measure the interfacial concentrations of c_{Ai} and c_{Bi} directly. However, a simplification can be made by assuming that these concentrations are at equilibrium:

$$N = K_A(c_A - c_A^*), \quad (7)$$

where

$$c_A^* = c_B/m' \quad (8)$$

and the overall mass transfer coefficient is

$$K_A = \left(\frac{1}{k_A} + \frac{1}{m'k_B} \right)^{-1}. \quad (9)$$

The value of K_A may depend primarily on k_A or on k_B , depending on whether m' is very large or very small. For example, if acetic acid (C) is being extracted from *n*-hexane (A) by water (B), the distribution ratio m' is very large and from Eq. (9) we see that $K_A \approx k_A$. In this case the extraction rate is controlled by the *n*-hexane phase resistance. Conversely, if the solute (e.g., benzoic acid) is much less soluble in the water than in the *n*-hexane, then $K_A = k_B/m'$ and the extraction rate is controlled by water phase resistance.

Mass transfer to or from droplet dispersions is employed in nearly all types of extraction equipment, so it is important to be able to estimate the two values of k for the droplet phase (dispersed) and the surrounding liquid phase (continuous).

In the absence of interfacial contamination, the motion of a droplet through surrounding liquid sets up toroidal circulation within the drop, and mass transfer coefficients are increased.

Surface-active contaminants, even in trace concentrations, tend to be adsorbed on the droplet surface and reduce or totally prevent internal circulation. This is particularly the case for smaller (<1 mm) droplets. For internally stagnant droplets, the droplet phase mass transfer coefficient is given by the approximate expression

$$k_d \approx 6.6D_d/d, \quad (10)$$

where d is the droplet diameter. In the continuous phase, the mass transfer coefficient expression is similar to that for a solid spherical particle moving through a fluid:

$$k_c d/D_c = 2 + 0.6(\rho_c u d/\mu_c)^{0.5} (\mu_c \rho_c/D_c)^{0.33}, \quad (11)$$

where ρ is the density, u the velocity, and μ the viscosity.

While interfacial contaminants tend to reduce the mass transfer coefficients by causing the droplets to be stagnant rather than circulating, another surface effect may enhance mass transfer. This is the Marangoni effect, whereby local variations in interfacial tension due to the mass transfer process itself can create rapid motions (interfacial turbulence) at the interface.

When chemical reactions occur in an extraction process, the effective mass transfer coefficient may be higher or lower than that expected from purely physical considerations, e.g., Eqs. (10) and (11). For example, the slow interfacial reaction of Eq. (5) will tend to reduce the mass transfer rate. On the other hand, a rapid irreversible reaction can enhance the mass transfer rate.

C. Interfacial Area and Droplet Behavior

As noted earlier, the rate of mass transfer is proportional to the interfacial area. Effective mass transfer requires that a high specific interfacial area (interfacial area per unit volume) should be created. For a droplet dispersion such as exists in most extraction processes, the specific interfacial area is given by

$$a = 6\Phi/d_{32}, \quad (12)$$

where Φ is the holdup (volume fraction) of the dispersed phase and d_{32} the Sauter mean droplet diameter calculated from the size distribution as follows:

$$d_{32} = \sum_{i=1}^n d_i^3 / \sum_{i=1}^n d_i^2. \quad (13)$$

In the absence of strong agitation, drop sizes are determined primarily by the conditions of formation and detachment at solid surfaces such as nozzles or pieces of packing.

Interfacial tension (γ) and buoyancy are of prime importance. Under conditions of highly turbulent agitation, the droplet size is determined by breakup with fluid eddies, and the relationship

$$d_{32} = K' \frac{\gamma^{0.6}}{\rho^{0.2} \Psi^{0.4}}, \quad (14)$$

where Ψ is the energy dissipation rate per unit volume, has been found useful. The dimensionless quantity K' varies between about 0.02 for stirred tanks (localized high turbulence) to 0.36 for reciprocating plate columns (uniformly distributed turbulence). Drop diameter can also be affected by coalescence, either between the drop and solid surface or between the drops themselves.

The holdup Φ is just as important as d_{32} in Eq. (12) for specific interfacial area. In a batch extraction it is simply determined by the amounts of each phase added, but in many types of continuous countercurrent contactor it is a complex function of flow rates, drop size, etc. The slip velocity u_s for countercurrent flow is defined in terms of the superficial velocities U (flow rates per unit cross-sectional area) of each phase,

$$\frac{U_d}{\Phi} + \frac{U_c}{1 - \Phi} = u_s = u_k(1 - \Phi). \quad (15)$$

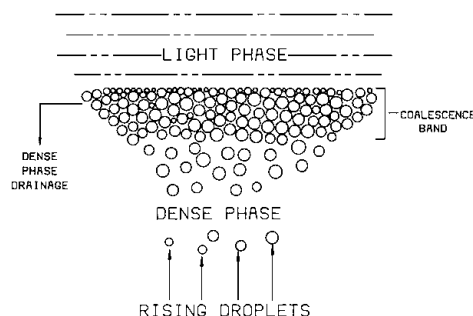


FIGURE 4 Coalescence of a droplet dispersion for the case of a denser continuous phase.

The characteristic velocity u_k is a function of droplet size, density difference, viscosity, etc. Thus, the holdup tends to increase either as the superficial flow velocities U_c and U_d are increased or as the characteristic velocity is reduced (e.g., by increasing agitation). A point is eventually reached where the increase in holdup becomes unstable (typically when $\Phi = 0.3$ – 0.4). This phenomenon is known as flooding, and it imposes a limit on the flow rates and agitation levels that can be used in countercurrent extraction processes.

In order that the two liquid phases can be drawn off separately after an extraction process, the droplets must be allowed to coalesce in the absence of agitation. Coalescence occurs by slow drainage of the continuous phase from the interstices between droplets (Fig. 4). When the films separating adjacent interfaces become thin enough for intermolecular forces to act, they rupture, causing rapid coalescence. Coalescence occurs readily in systems having a high interfacial tension and density difference, low viscosity, and absence of interfacial contaminants. For systems that are slow to coalesce, various promoters have been developed. A popular aid is a knitted mesh of stainless steel and polypropylene which enhances coalescence of either aqueous droplets (on the steel surfaces) or organic droplets (on the polypropylene surfaces). Coalescence can also be accelerated by applying an alternating voltage across the continuous phase, but this technique is only applicable if the continuous phase is nonaqueous.

Complete coalescence of droplets can be difficult due to the formation of a fine haze of secondary droplets. Although this haze may comprise less than 1% of the process flow, it is often an unacceptable loss. Another problem that can arise in continuous coalescers is the accumulation of interfacial contaminant ("crud"). This necessitates periodic purging of the coalescence area with fresh solvent.

D. Stagewise Processes

A single equilibrium stage (Fig. 1) is seldom sufficient to achieve the desired extraction factor. Thus, in most

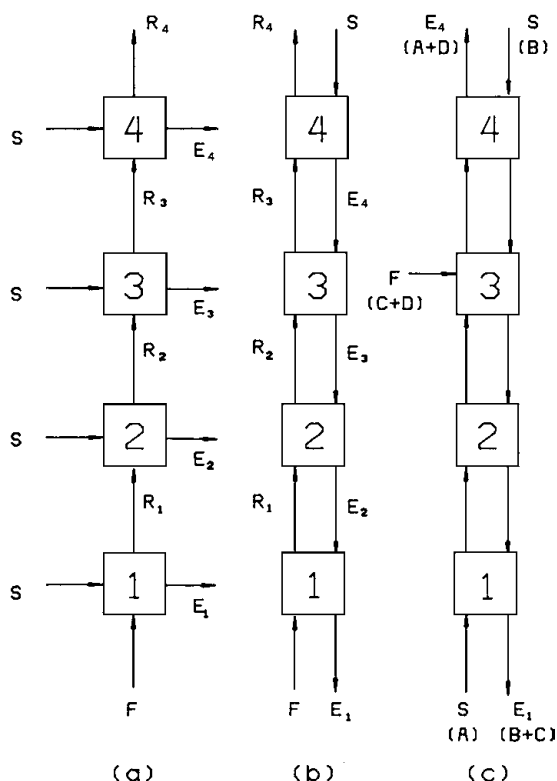


FIGURE 5 Multistage extraction processes: (a) crosscurrent extraction, (b) countercurrent extraction, and (c) countercurrent fractional extraction.

extraction processes, an arrangement of many stages is necessary. Three possible schemes are shown in Fig. 5.

Crosscurrent extraction (Fig. 5a) consists of repeated contacts of the feed solution with fresh solvent, resulting in a series of extract streams of gradually diminishing concentration. It is a simple arrangement and can be readily applied batchwise (as a laboratory operation) or with continuous flow of feed and solvent. It is more effective than a single-stage fed with the same total flows of solvent and feed.

Countercurrent extraction (Fig. 5b) is more effective than crosscurrent extraction because all the fresh solvent is contacted with the weakest raffinate. This arrangement is almost universally used in industrial extraction equipment.

Fractional countercurrent extraction (Fig. 5c) is commonly used to separate a mixture of two solutes (C and D) using two solvents (A and B). In this case the feed is introduced at a stage within the countercurrent arrangement, but the principle is the same as in the single-solute scheme shown in Fig. 5b.

The important question for the designer of a continuous countercurrent extraction process is: how many equilibrium stages are needed, given the flow rates and inlet and outlet concentrations? A simple graphical procedure is possible for a single-solute extraction when the two

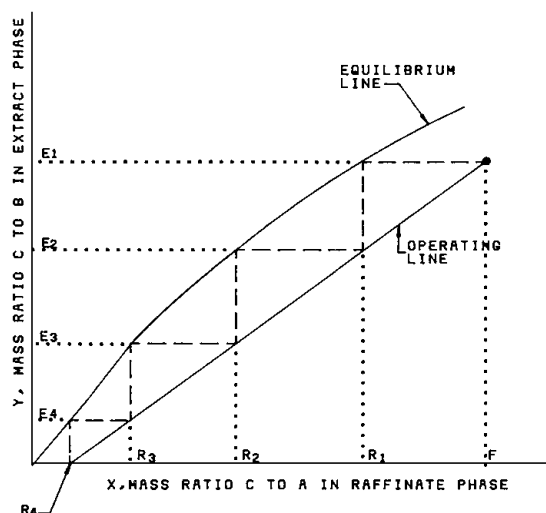


FIGURE 6 Estimation of the number of countercurrent stages for the immiscible solvent case.

solvents A and B can be considered to be immiscible; this is summarized in Fig. 6. The straight “operating line” represents the actual compositions of each phase at different interstage levels in the countercurrent arrangement as shown. The “equilibrium line,” usually curved, represents the compositions of each phase at equilibrium. In carrying out the stepwise construction (dashed lines) it is assumed that equilibrium is reached between the streams leaving each stage (R_1 and E_1 , etc.). The number of steps that must be drawn between the given compositions at each end of the cascade is the number of countercurrent stages required.

This type of design procedure is based on two additional assumptions which may not always be true in practice. First, it is assumed that equilibrium is completely attained in each stage. In practice this is usually not the case because mass transfer is a first-order rate process [Eq. (8)] and complete equilibrium is only reached asymptotically. A second factor is that in many types of continuous countercurrent equipment, some reverse flow (backmixing) occurs; for example in Fig. 5b, a small portion of stream R_2 may find its way back into stage 1.

These two effects are major factors for the need to define an overall stage efficiency, which is the ratio of the number of ideal stages theoretically required for a given separation to the number actually required.

E. Differential Contact

Many types of countercurrent equipment (e.g., packed columns) do not contain discrete compartments and therefore cannot be treated as combinations of discrete stages (Fig. 5b). The term “differential contactor” is used to describe this category, and a somewhat different design

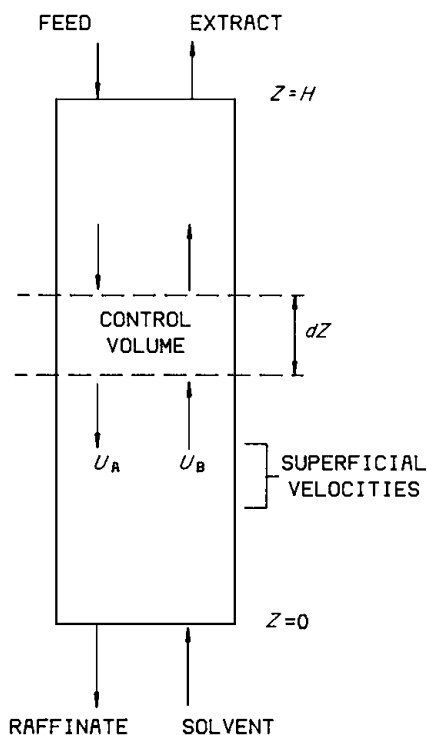


FIGURE 7 Control volume for analysis of countercurrent differential extractor.

approach must be used. The design basis is shown in Fig. 7, in which a differential control volume dz (for unit cross-sectional area of column) is defined. A material balance equation for the control volume equates the amount of solute transferred between phases with the mass transfer rate expression including the effects of concentration driving force [Eq. (9)] and specific interfacial area,

$$U_A dc_A = U_B Dc_B = aK_A(c_A - c_A^*) dz. \quad (16)$$

Integration of this equation between $z = 0$ and $z = H$ gives

$$H = \left(\frac{U_A}{aK_A} \right) \int_{C_{A0}}^{C_{AH}} \frac{dc_A}{C_A^* - C_A}. \quad (17)$$

The integral can be found from the operating and equilibrium lines (Fig. 6) for the process and is known as the number of transfer units (NTU). It is a unitless measure of the degree of separation required. The term in parentheses in Eq. (17) is the height of a transfer unit (HTU), which is a characteristic of the flow conditions and mass transfer effectiveness of the contactor. The more effective conditions correspond to smaller values of the HTU whereby a shorter contactor is required for a given NTU. If the equilibrium line and the operating line are approximately parallel, the NTU is approximately equivalent to the number of ideal stages required.

Although the transfer unit concept allows for the effects of mass transfer rate, in the simplified form given here it

does not allow for backmixing effects. Therefore contactor lengths calculated by Eq. (17) will tend to be too low if backmixing is significant.

The rate of transport within a phase due to backmixing can be described by an equation analogous to the diffusion equation:

$$N = -E(dc/dz). \quad (18)$$

In this case, E is the axial dispersion coefficient and z refers to the axial direction in the contactor. The axial dispersion coefficient E is a function of flow rates, turbulence, etc., and has a value far in excess of the molecular diffusivity D . Design methods that allow for axial dispersion are described in the research literature, but there is an acute need for more data on values of E for large-scale equipment.

II. INDUSTRIAL EXTRACTION EQUIPMENT

Industrial application of solvent extraction has increased rapidly over the past several decades. Probably more types of extractors for solvent extraction have been developed than for any other chemical engineering unit operation. They have also been developed for specific processes with which they tend to become associated. As a result, selection of extractors can be quite bewildering to someone who is contemplating a new process application. Choosing an extractor is still an art as well as a science. The following criteria should be considered in the selection of an extractor for a particular application: stability and residence time, settling characteristics of the solvent system, number of stages required, capital cost and maintenance, available space and building height, throughput, and reliability of operation. These factors should be evaluated at an early stage before the pilot-plant tests are carried out. Although cost ought to be a major balancing consideration, in many actual cases previous experience and practice are the deciding factors.

Pilot-scale testing remains an inevitable preliminary to a full-scale extractor design for any new commercial process. The pilot-scale extractor should be of the same type as the full-scale extractor. In the present stage of knowledge, reliable scale-up to industrial-scale extractors still depends on the correlations based on extensive performance data collected from pilot-scale and large-scale extractors covering a wide range of liquid systems.

Commercial extractors can be classified according to the methods applied for interdispersing phases and producing the countercurrent flow pattern. Both of these can be achieved either by the force of gravity acting on the density difference between phases or by applying centrifugal force.

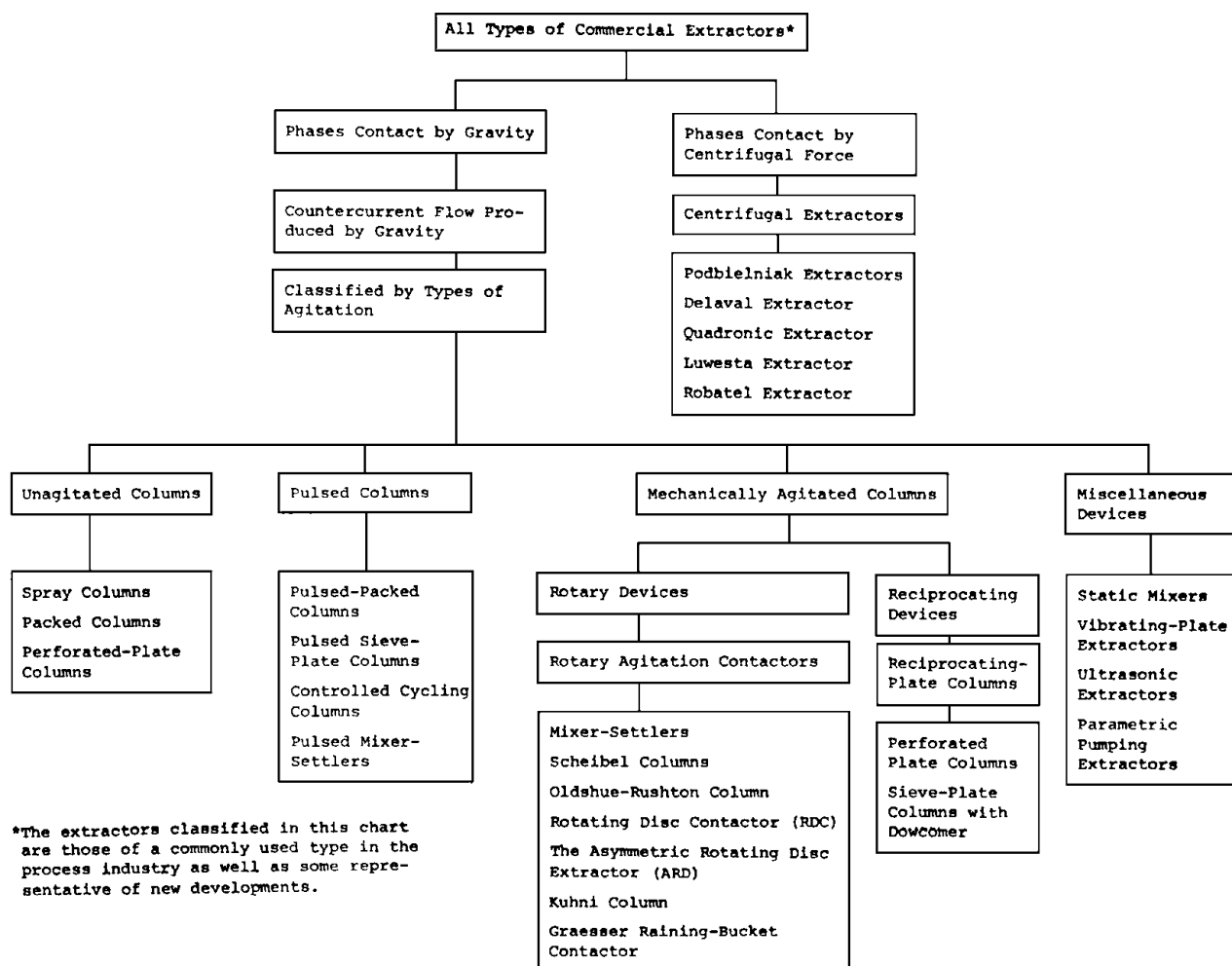


FIGURE 8 Classification of commercial extractors. [Reprinted with permission from "Kirk-Orthmer Encyclopedia of Chemical Technology," 4th ed., Vol. 10 (1994), Wiley (Interscience), New York. © 1994 John Wiley & Sons, Inc.]

Figure 8 summarizes the classification of major types of industrial extractors, while Table I summarizes their features and fields of industrial application.

A. Unagitated Columns

Unagitated columns are the simplest possible extraction columns from a mechanical construction point of view, and three types are shown in Fig. 9. Spray columns consist of a device to spray the dispersed phase through a column of the continuous phase. Either light or heavy phase can be dispersed. Spray columns generally provide one or, at most, two equilibrium stages. Because of their simple construction, they are used in industry for operations such as washing and neutralization, which often require no more than one or two stages. Packed columns provide better mass transfer efficiency because the packing promotes mass transfer and reduces backmixing. It is important that

the packing should be preferentially wetted by the continuous phase to avoid coalescence of the dispersed phase, which reduces the interfacial area per unit volume. To reduce the effects of channeling, redistribution of the liquids at fixed intervals is normally required in the taller columns. A high-efficiency packing-SMR (super mini ring or QH-1) has been developed in China for liquid-liquid extraction. Columns up to 2.6 m in diameter, used in oil refining and wastewater treatment, have been reported.

Perforated-plate columns are semistage in operation. They are reasonably flexible and efficient. A 7-ft-diameter, 80-ft-high perforated-plate column used for extraction of aromatics was reported to have the equivalent of 10 theoretical stages. Because of the simplicity and low cost of packed and perforated-plate columns, they are widely used in industry despite their low efficiency, particularly for processes requiring few theoretical stages and for corrosive systems where absence of mechanical moving parts

TABLE I Summary of Features and Fields of Industrial Application of Commercial Extractors

Types of extractor	General features	Fields of industrial application
Unagitated columns	Low capital cost, low operating and maintenance cost, simplicity in construction, handles corrosive material	Petrochemical, chemical
Mixer-settlers	High-stage efficiency, handles wide solvent ratios, high capacity, good flexibility, reliable scale-up, handles liquids with high viscosity	Petrochemical, nuclear, fertilizer, metallurgical
Pulsed columns	Low HETS, no internal moving parts, many stages possible	Nuclear, petrochemical, metallurgical
Rotary agitation columns	Reasonable capacity, reasonable HETS, many stages possible, reasonable construction cost, low operating and maintenance cost	Petrochemical, metallurgical, pharmaceutical, fertilizer
Reciprocating-plate columns	High throughput, low HETS, great versatility and flexibility, simplicity in construction, handles liquids containing suspended solids, handles mixtures with emulsifying tendencies	Pharmaceutical, petrochemical, metallurgical, chemical
Centrifugal extractors	Short contacting time for unstable material, limited space required, handles easily emulsified material, handles systems with little liquid density difference	Pharmaceutical, nuclear, petrochemical

is advantageous. A 40-ft-diameter perforated-plate column with downcomers between each stage has been used for petroleum processing.

B. Mixer-Settlers

Mixer-settlers are widely used in the chemical process industry because of their reliability, flexibility, and high

capacity. They are particularly economical for operations that require high capacity and few stages. Mixer-settlers with a capacity of up to 6000 gal/min have been used in the hydrometallurgical extraction of copper. The main disadvantages of mixer-settlers are their size, particularly the large ground space requirement, and the inventory of material held up in the equipment. In the past decades considerable developmental work has been done

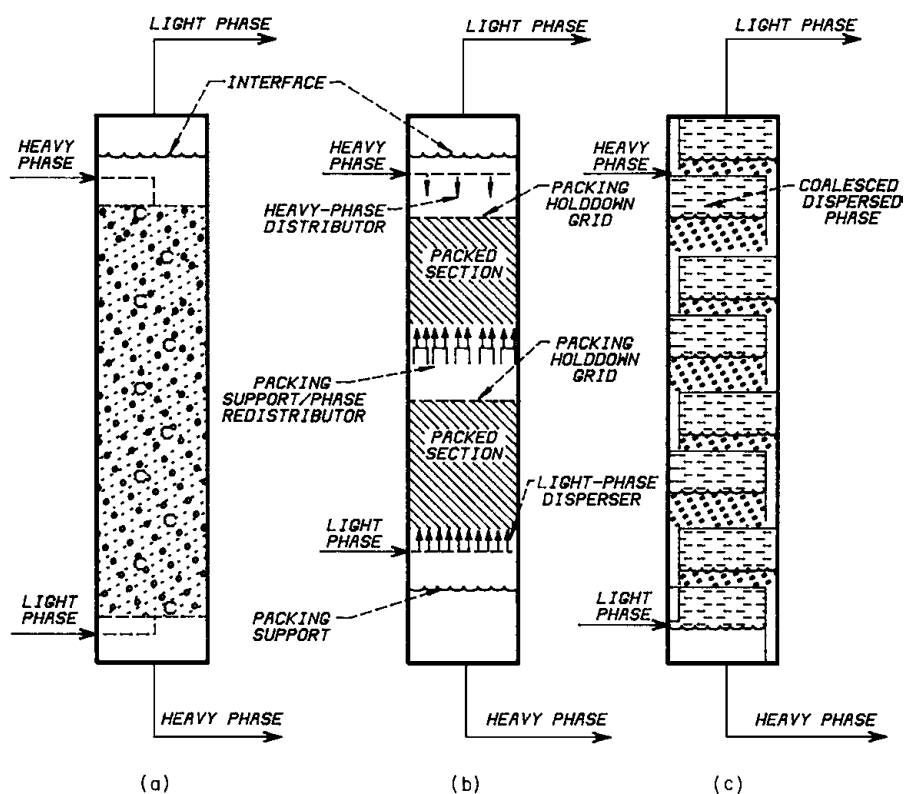


FIGURE 9 Unagitated column extractors: (a) spray, (b) packed, and (c) sieve tray.

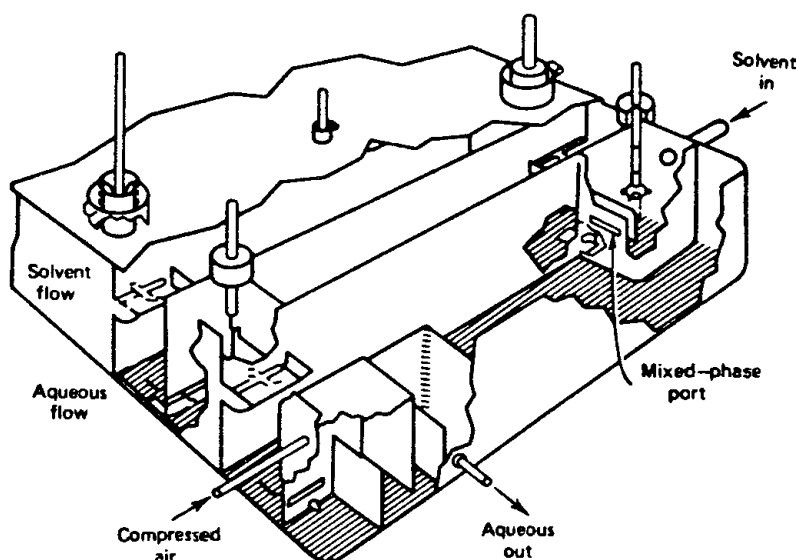


FIGURE 10 Box-type mixer-settler. (Courtesy of the Institution of Chemical Engineers.)

to improve the contactors, and many new devices have been reported.

The simple box-type mixer-settler (Fig. 10) developed by the British Atomic Energy Authority has been used extensively in the separation and purification of uranium and plutonium. The design avoids all interstage piping by use of a partitioned box construction and involves no interstage pumping. The driving force for the flow is derived simply from the density difference between the stages. The Davy-McKee mixer-settler (Fig. 11a) is a type of mixer-settler with the pump-mix approach. The liquids run through a draft tube and are mixed and pumped by an impeller running directly above the draft tube. The dispersed phases flow off the top of the mixer and down through a channel into a rectangular settler. The recently developed Davy-McKee combined mixer-settler (CMS) (Fig. 11d) consists of a single vessel in which, under operating conditions, three zones coexist. These are (i) an upper separated organic-phase zone, (ii) a lower separated aqueous-phase zone, and (iii) a central zone containing mixed phases or dispersion. Commercial units used for uranium recovery have been reported. The IMI pump-mixer-settler (Fig. 11b) has been widely used in many process industries. A unit with a capacity of 2200 gal/min has been used in phosphoric acid plants. The pumping device is not required to act as the mixer, and the two phases are dispersed by a separate impeller mounted on a shaft running coaxially with the drive to the pump. The General Mills mixer-settler (Fig. 11c) is a type of pump-mix unit designed for metallurgical extraction. The unit has a baffled cylindrical mixer fitted in the base with a turbine which does both mixing and pumping for the incoming liquids. The dispersion leaves from the top of the mixer

and flows into a shallow rectangular settler designed for minimum holdup.

There are two distinct types of mixer-settlers developed by the Lurgi Company in Germany. The horizontal type has an axial-flow impeller to both mix and pump the phases. The other type has a vertical stack of stages and takes the form of a column. Mixing and phase transfer take place in pumps attached to the sides of the settling column, and phases flow interstage through a complex arrangement of baffles within the settling zones. Since the contact time in each mixing stage is short, this extractor is more suitable for processes in which the rate of mass transfer is fast. The units, which have up to 1800 ton/hr capacity, have been commercially proven. Columns up to 10 ft in diameter have been used for an *N*-methylpyrrolidone (NMP) aromatic extraction process. Holmes and Narver has marketed a new type of mixer-settler which incorporates multicompartment mixers. The unit has been used in hydrometallurgical extractions for copper refining. Motionless in-line mixers utilize the energy due to the flow and the pressure drops for mixing and dispersing the phases. Performance data on static mixers and on Sulzer mixers have been reported.

Mixer-settlers are relatively reliable on scale-up because they are practically free of interstage backmixing and stage efficiencies are high. Mixers can be scaled up 200-fold on throughput by geometric similitude at constant power input per unit mixer volume. The flow capacity of settlers is determined by the band of dispersion at the interface, the thickness of which is a measure of the approach of flooding. The thickness of the band increases exponentially with increasing flow per unit area, and settlers can be scaled up by factors of up to 1000 on this basis.

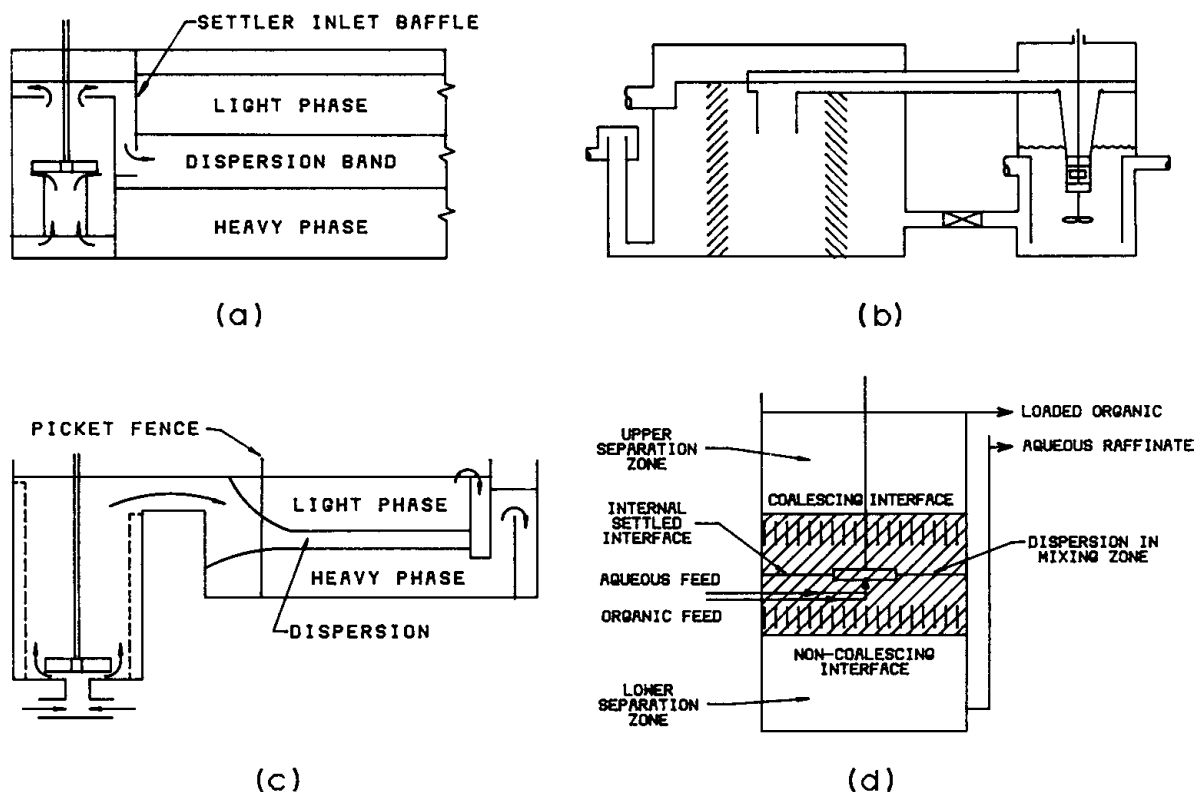


FIGURE 11 Typical mixer-settlers. (a) Davy-McKee, (b) IMI, (c) General Mills, and (d) CMS.

In large industrial mixer-settlers, the settlers usually represent at least 75% of the total volume of the units.

C. Pulsed Columns

The application of mechanical pulsation of the liquid increases both turbulence and interfacial areas and greatly improves the mass transfer efficiency compared with an unpulsed column. Axial mixing in a pulsed column is relatively small compared with mechanical rotary-agitated columns. This leads to a substantial reduction in HETS or HTU values. The pulsed packed column consists of a vertical cylindrical vessel filled with packing. Light and heavy liquids, either of which is dispersed in the form of drops, pass countercurrently through the column and are simultaneously moved up and down by means of a pulsating device connected to the bottom of the column through a side-entering "pulse leg." Mechanical difficulties with the generation of the pulse formerly limited pulsed columns to comparatively small diameters, but pulsed packed columns up to 9 ft in diameter using a rotary pulsing device have been reported. Generation of pulsations by compressed air has received increasing attention.

Pulsed perforated-plate columns (Fig. 12) are fitted with horizontal perforated plates or sieve plates which occupy the entire cross section of the column. The total free area of

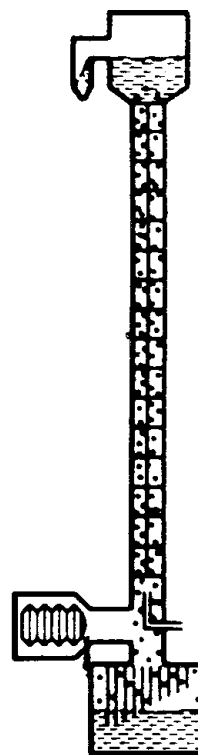


FIGURE 12 Pulsed-perforated-plate column. (Reprinted courtesy of ERIES, France.)

the plate is typically 20–25%. The columns are generally operated at frequencies of 1.5–4 Hz with amplitudes of $\frac{1}{4}$ –1 in.

Uniform distribution of energy over a cross section of the column and hence uniform distribution of drops in the column are the unique features of pulsed perforated-plate columns. They give low axial mixing and high extraction efficiency. Pulsed perforated-plate columns up to 3 ft in diameter have been widely used in the nuclear industry, and even larger columns are being developed in the nuclear fuel reprocessing industry because of the increasing demand for nuclear power. A new type pulsating plate (with Guiding vanes) has been developed by I. J. Gorodetsky *et al.* in the former Soviet Union. Columns up to 2 m in diameter have been used in industry.

D. Mechanically Agitated Columns

Mechanically agitated columns (Fig. 13) can be divided into two main classes according to the type of agitation provided: rotary-agitated columns and reciprocating or vibrating-plate columns. Rotary agitation provides many mechanical advantages. Most modern differential contactors employ this method. Various types of commercial rotary-agitated columns are listed in Fig. 8. Of these, the Scheibel column, the rotating disk contactor, the Oldshue–Rushton multiple-mixer column, and the Kuhni column are types of rotary agitated columns that have been proven in industrial installations.

There are several design variations of the Scheibel column (Fig. 13a). The earlier model consists of alternate compartments agitated with impellers, and the oth-

ers are packed with an open woven wire mesh. A newer type of column using horizontal baffles was developed in 1956. It improves the HETS and permits a more efficient scale-up to large-diameter columns. There are two configurations of design; one is with wire mesh packing and one is without packing. Performance data for a 12-in. column with and without wire mesh packing have shown that the HETS of this type of column varies as the square root of the diameter. A third design is basically similar to the second design, but a pumping impeller instead of a turbine impeller is used in the mixing stage. Scheibel columns up to 8.5 ft in diameter are in service.

The rotating disk contactor (RDC) (Fig. 13b), developed in Europe by Reman in 1951, uses the shearing action of a rapidly rotating disk to interdisperse the phases. Rotating disk contactors have been widely used throughout the world, particularly in the petroleum and petrochemical industries, where they have become associated with furfural and SO₂ extraction, propane deasphalting, sulfolane extraction for separation of aromatics from aliphatics, lube oil, and carpolactam purification. Columns up to 14 ft in diameter are in service. The RDC design is probably the best known because of wide application and extensive experimental performance studies. The extensive study reported by Strand *et al.* has provided an excellent theoretical framework for the scale-up of RDCs. Misek has done excellent work on compiling a design manual based on his own work and literature results.

The asymmetric rotating disk (ARD) contactor (Fig. 13c) was developed by Misek and coworkers in Czechoslovakia and has been increasingly used in western

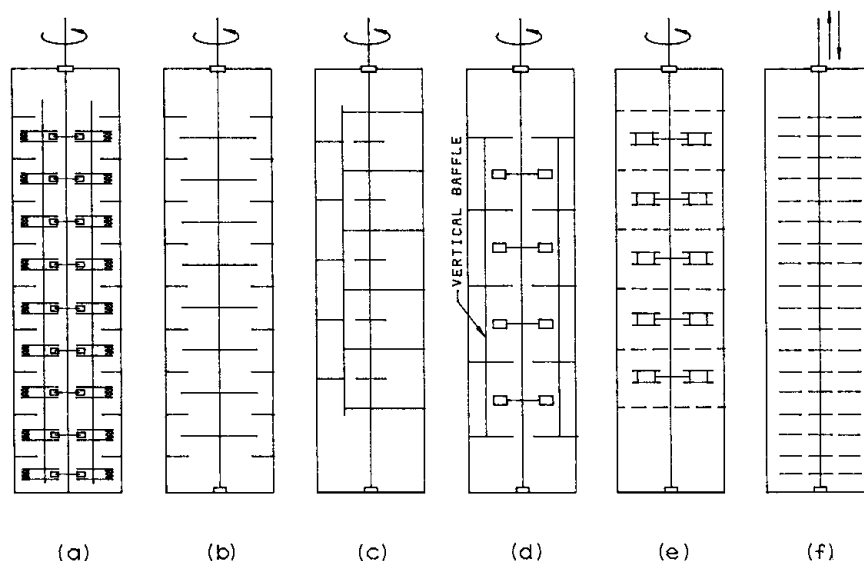


FIGURE 13 Typical mechanically agitated columns. (a) Scheibel, (b) RDC, (c) ARD, (d) Oldshue–Rushton, (e) Kuhni, and (f) reciprocating-plate.

Europe. Its design is aimed at retaining the efficient shearing action of the RDC by using the rotating disk to produce dispersion and to reduce backmixing by means of the coalescence–redispersion cycle produced in the separated transfer settling zones. The ARD extractor is used for extraction of petrochemicals, pharmaceuticals, and caprolactam, as well as for propane deasphalting, phenol removal from wastewater, furfural refining of oils, etc. Extensive research work has been done and performance data in large-scale extractors have been accumulated by Misek *et al.* to provide basic information for scale-up and design.

The Oldshue–Rushton column (Fig. 13d) was also developed in the early 1950s and has been widely used in the chemical industry. It consists essentially of a number of compartments separated by horizontal stator-ring baffles, each fitted with vertical baffles and a turbine-type impeller mounted on a central shaft. Columns up to 9 ft in diameter have been reported in service.

The Kuhni contactor (Fig. 13e) has gained considerable commercial application in Europe. Its principal features are the use of a shrouded turbine impeller to promote radial discharge within the compartments and a variable hole arrangement to allow flexibility of design for different process applications. Kuhni extractors are used for extraction of petrochemicals and chemicals, phosphoric acid purification, as well as hydrometallurgical applications and wastewater treatment. Columns up to 16.5 ft in diameter have been constructed.

The RTL contactor, formerly known as the Graesser raining bucket contactor (Fig. 14), is a horizontal design with the phases interdispersed by “water wheel” arrangements. The unit has the unusual feature of dispersing each phase into the other. The contactor was developed for

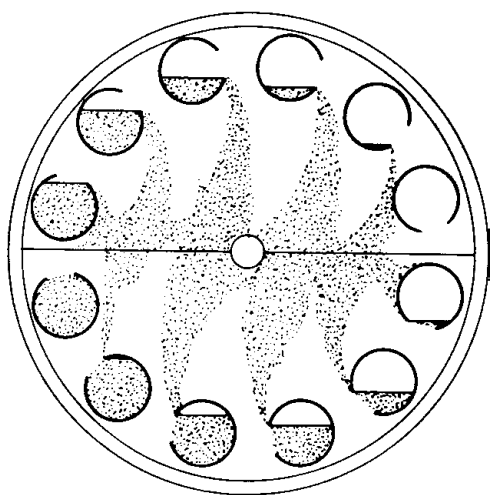


FIGURE 14 Graesser raining-bucket contactor. [Reprinted by permission of *Chem. Eng.* **75**(18), 76 (1968).]

handling the difficult settling systems found in the coal tar industry, and it has proved attractive for other applications as well. It is suitable for handling solid–liquid systems. Units have been built from 4 in. (100 mm) to 6 ft (1.8 m) in diameter.

E. Reciprocating-Plate Columns

In 1935, Van Dijk proposed that the extraction efficiency of a perforated-plate column could be improved by either pulsing the liquid contents of the column or by reciprocating the plates. The latter idea was relatively little exploited until the late 1950s, when Karr reported data on a 3-in.-diameter, open-type, perforated reciprocating-plate column. The column was further developed by Karr and Lo by employing baffles.

The open-type perforated reciprocating-plate column (Fig. 13f) developed by Karr and Lo consists of a stack of perforated plates and baffle plates which have a free area of about 58%. The central shaft which supports the plates is reciprocated by means of a reciprocating drive mechanism located at the top of the column. These columns have gained increasing industrial application in the pharmaceutical, petrochemical, and wastewater-treatment industries, and columns up to 60 in. in diameter are in service.

Prochazka *et al.* reported two new types of reciprocating-plate columns. One type of column uses perforated plates with segmental passages for the continuous phase, with the dispersed phase passing through the relatively small perforations in the plates. Because of the segmental passages for the continuous phase, the throughput of the column is reported to be relatively higher than that of pulsed or other types of extractors. In the second design, the plates are carried by two shafts alternately attached to one and free to slide on the other. The two interlaced sets of plates are thus, reciprocated 180° out of phase. Commercial applications of this type of column up to 47 in. in diameter have been reported in Eastern Europe and Russia.

The KRIMZ and GIAP types of RPC were developed in the former USSR and the plate designs feature rectangular punched perforations where the displaced metal strip remains attached as inclined vanes. The purpose of the vanes is to deflect the liquid and give it radial motion, which, can be beneficial in reducing axial mixing in large-diameter columns. The column have gained increasing industrial applications in the petrochemical, pharmaceutical, and wastewater-treatment industries. Columns of up to 1.6 m diameter have been installed for caprolactam extraction in the former USSR and Eastern Europe.

F. Centrifugal Extractors

Centrifugal extractors offer short contact times and increase the efficiency of phase separation by application

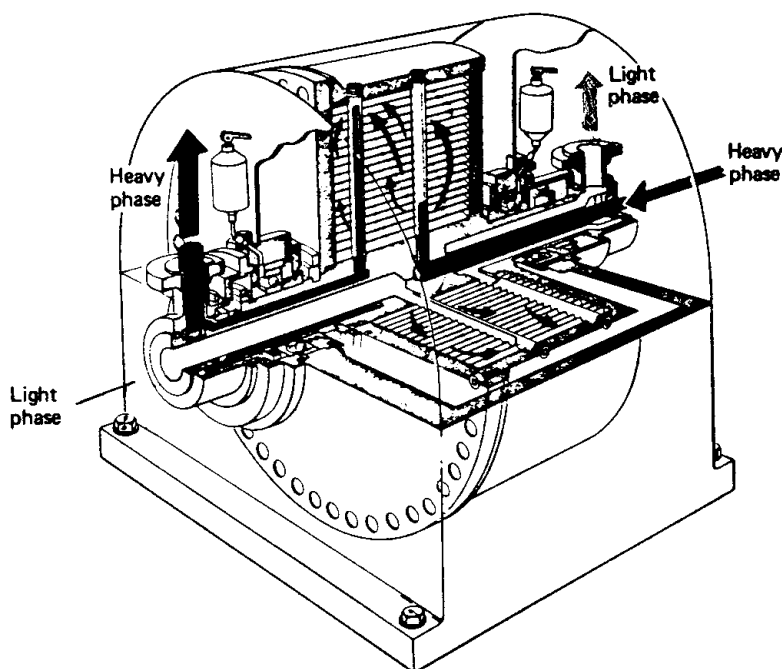


FIGURE 15 Podbielniak centrifugal extractor. (Courtesy of Baker Perkins, Inc.)

of centrifugal force. The units are compact, and relatively high throughput can be achieved in a small volume. They have been favored for applications involving either chemically unstable (e.g., extraction of antibiotics) or slow-settling systems.

The Podbielniak extractor (Fig. 15) was the first centrifugal unit introduced to industry in the early 1950s and is still probably the best known. Its design consists essentially of a perforated-plate extraction column that has been wrapped around a shaft, which, in turn, is rotated to create a centrifugal force field that achieves a great reduction in the height and contacting time of a perforated-plate column. The extractors have been widely used in the pharmaceutical industry (e.g., extraction of penicillin) and are increasingly used in other fields as well. Commercial units with throughput up to 26,000 gal/hr have been reported.

It is claimed that the Alfa-Laval extractor can give up to 20 theoretical stages in one unit. The capacity of the standard unit depends on the systems being handled and ranges from 1500 to 5600 gal/hr. Antibiotic extractions and petrochemical processing are typical applications.

The Westfalia centrifugal extractor is built on a vertical rotating principle and is available with up to three contact stages. One advantage is that the light phase does not have to be introduced under pressure. It has been reported that the capacity of the largest model ranges from 2000 gal/hr with three stages to 13,000 gal/hr with a single stage.

The Robatel extractor is a series of mixer-settlers stacked on their sides, with mixing in each stage being carried out by a stationary disk attached to the shaft while

the mixing chamber revolves. The two phases pass into the separating chamber, which contains no coalescing plates, and then pass via a channel system into adjacent stages in countercurrent flow. The unit generally provides three to eight stages, and throughputs up to 1600 gal/hr have been reported. The Robatel extractors have been extensively used in the nuclear industry.

A new countercurrent continuous centrifugal extractor, developed in the former USSR, has the feature that mechanical seals are replaced by liquid seals with the result that operation and maintenance are simplified; the mechanical seals are an operating weak point in most centrifugal extractors. The operating units range between 400 and 1200 mm in diameter, and a capacity of 70 m/hr has been reported in service. The extractors have been applied in coke-oven refining, erythromycin production, lube oil refining, etc.

Research and development on other nondispersive forms of contactors, e.g., Hi-Gee solvent extractors that give a high efficiency per unit volume, and contactors effective with very short residence times, e.g., improvement on the centrifugal extractor, has been reported.

III. INDUSTRIAL EXTRACTION PROCESSES

A. Organic Processes

The industrial applications of solvent extraction to organic processes are very extensive, and only those processes of major industrial importance are briefly reviewed here.

1. Petroleum and Petrochemical Processes

The first large-scale application of solvent extraction was the removal of aromatics from kerosene to improve its burning properties. Solvent extraction is used for processing jet fuel and lubricating oil, which require a low aromatic content. Solvent extraction is used equally extensively to meet the ever-increasing demand for high-purity aromatics such as benzene, toluene, and xylene (BTX) as feedstocks for the petrochemical industry. The separation of aromatics from aliphatics is one of the largest applications of solvent extraction.

a. Lubricating oil extraction. Aromatics are removed from lubricating oils to improve the viscosity index and chemical stability. The solvents used in the commercial processes for the extraction of lubricating oil are furfural, phenol, and liquid sulfur dioxide. Each of these solvents has a high capacity for dissolving hydrocarbons and a particular selectivity for aromatic rather than aliphatic hydrocarbons. Among the solvents, most new plants are adopting furfural processes. A process known as EXOL N, using *N*-methyl-2-pyrrolidone (NMP) as solvent, is employed internationally in at least a dozen full-scale lube extraction plants. A useful comparison of the various processes is available.

b. Separation of aromatic and aliphatic hydrocarbons. Aromatics extraction for aromatics production, jet fuel, kerosene treatment, and enrichment of gasoline fractions is one of the most important applications of solvent extraction. The features of various commercial processes are summarized in Table II. The sulpholane process (Fig. 16), introduced by the Shell Company in 1962, has become one of the most important advances in large-scale aromatics production. Sulpholane, thiocyclopentane 1,1-dioxide ($(\text{CH}_2)_4\text{SO}_2$), is a strong polar compound that is highly selective for aromatic hydrocarbons. It also has a much greater solvent capacity for hydrocarbons than glycol systems. Additional features are its high density, low heat capacity, and good stability. The design of the sulpholane process usually includes the rotating disk contactor, which is also a Shell development. Figure 16 shows a schematic diagram of the sulpholane process. Many large units have been built all over the world, and there is little doubt that the use of this process will continue to increase.

c. Butadiene. Solvent extraction is used in the separation of butadiene from other C_4 hydrocarbons in the manufacture of synthetic rubber. The butadiene is produced by catalytic dehydrogenation of butylene and the liquid butadiene is then extracted with an aqueous cuprammonium acetate solution with which the butadiene re-

acts to form a complex. Butadiene is then recovered by stripping from the extract. Distillation is a competing process.

d. Caprolactam extraction. Caprolactam is the monomer for Nylon 6. The purification of caprolactam is very important since fiber-grade caprolactam requires extremely high purity. Solvent extraction has been used for purification of the crude aqueous caprolactam. Aromatic hydrocarbons such as toluene are used as the solvent. A detailed description of the process is available.

e. Anhydrous acetic acid. In one process for the manufacture of acetic acid by direct oxidation of a petroleum-based feedstock, solvent extraction has been used to separate acetic acid from aqueous acid reaction liquor containing significant quantities of formic and propionic acids. The process developed by the Distiller's Company uses isoamyl acetate as a solvent to extract the aqueous feed to remove nearly all of the acetic acid with some of the water. The extract is then dehydrated by azeotropic distillation using the same organic solvent as a water entrainer. The aqueous raffinate from the extractor contains small quantities of acid and solvent which are recovered by steam stripping before the water is discarded. It is claimed that the extraction step in this process gives substantial savings in both plant capital investment and in operating cost.

f. Synthetic fuel. Solvent extraction has many applications in synthetic fuel technology. The extraction of Athabasca tar sands and Irish peat with *n*-pentane has been reported. A process for treating coal under hydrogen with a solvent has been described. The main object of the study was to open up coal with a minimum of hydrogen so that solvent extraction can be used to extract valuable feedstock components before the coal is burned. Solvent extraction is also used in coal liquification processes and in synthesis fuel refining.

2. Pharmaceutical Processes

Solvent extraction has been used extensively in the pharmaceutical industry because many pharmaceutical intermediates and products are heat sensitive and cannot be processed by methods such as distillation, etc. However, few details of current commercial operations have been published.

a. Antibiotics. Solvent extraction is an important step in the recovery of many antibiotics, such as penicillin, streptomycin, novobiocin, bacitracin, erythromycin, and the cephalosporins. The extraction of antibiotics can

TABLE II Solvents for the Separation of Benzene–Toluene–Xylene Mixtures from Light Feedstocks^a

Solvent	Process	Solvent additives and reflux conditions	Operating temperature	Contacting equipment	Comments
Sulfolane	SHELL process, Licensors: Universal Oil Products	Sulfolane selectivity and capacity insensitive to water content caused by steam-stripping during solvent recovery; heavy paraffinic countersolvent used	120°C	Rotating-disk contactor, up to 4 m in diameter	The high selectivity and capacity of sulfolane leads to low solvent/feed ratios, and thus smaller equipment
Glycol/water mixtures	UDEX process, Universal Oil Products	Solvent can be diethylene glycol and water, or a mixture of diethylene and dipropylene glycols and water, or tetraethylene glycol and water; light hydrocarbon reflux	150°C for diethylene glycol and water	Sieve-tray/extractor	Tetraethylene glycol and water mixtures are claimed to increase capacity by a factor of four and also require no antifoaming agent; the extract requires a two-step distillation to recover BTX
Tetraethylene glycol	Union Carbide Corp.	The solvent is free of water; a dodecane reflux is used which is later recovered by distillation	100°C	Reciprocating-plate extractor	The extract leaving the primary extractor is essentially free of feed aliphatics, and no further purification is necessary; two-stage extraction uses dodecane as a displacement solvent in the second stage
Dimethyl sulfoxide (DMSO)	Institut Français du Pétrole	Solvent contains up to 2% water to improve selectivity; reflux consists of aromatics and paraffins	Ambient	Rotating-blade extractor, typically 10–12 stages	Low corrosion allows use of carbon steel equipment; solvent has a low freezing point and is nontoxic; two-stage extraction has displacement solvent in the second stage
N-Methyl pyrrolidone (NMP)	AROSOLVAN process, Lurgi	A polar mixing component, either water (12–20% by weight) or monoethylene glycol (40–50% by weight) must be added to the NMP to increase the selectivity and to decrease the boiling point of the solvent; the NMP/water processes use a pentane countersolvent	NMP/glycol, 60°C; NMP/water, 35°C	Vertical multistage mixer-settler, 24–30 stages, up to 8 m in diameter	The quantity of mixing component required depends on the aromatics content of the feed
N-Formylmorpholine (FM)	FORMEX process, Snamprogetti	Water is added to the FM to increase its selectivity and also to avoid high reboiler temperatures during solvent recovery by distillation	40°C	Perforated-tray extractor, FM density at 1.15 aids phase separation	Low corrosion allows use of carbon steel equipment

^a From *Chem. Eng.* **83**(2), 86 (1976).

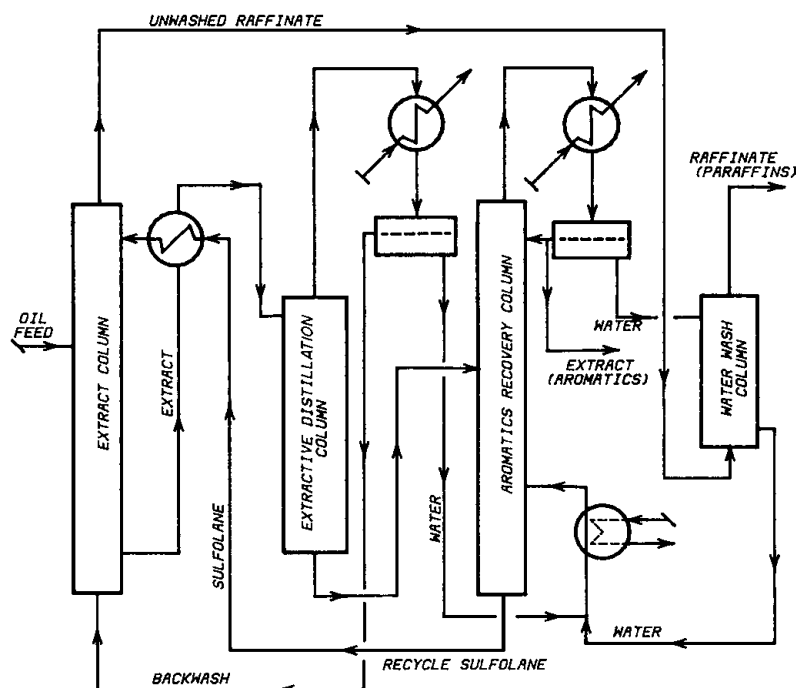


FIGURE 16 Sulfolane process for aromatic separation.

be typified by the production of penicillin, which is well documented. Penicillins are manufactured by batch fermentation. After filtration to remove mycelium, the aqueous fermentor broth is fed to a series of extractors for concentration and purification. In commercial practice, either amyl acetate or *n*-butyl acetate is used as the extraction solvent. The penicillin is first extracted into the solvent from the broth at a pH of 2.0–2.5. The extract is then treated with a buffer solution at a pH of approximately 6 to obtain a penicillin-rich aqueous solution. Finally, the pH is again adjusted to low value, and the penicillin is re-extracted into the solvent to yield a pure concentrated solution. The extraction operation is complicated by the fact that the penicillin degrades rapidly as the pH is reduced, and it is necessary to perform the extraction at low pH as quickly as possible. Centrifugal extractors are generally used to ensure a short residence time of the product. Supercritical extraction of pleuromutilin from fermentation using carbon dioxide has been reported.

b. Natural and synthetic vitamins. Solvent extraction is used extensively for preparation of natural and synthetic vitamins that are heat sensitive. Natural vitamins A and D are extracted from fish liver oils, and vitamin E from vegetable oils; liquid propane is the solvent. In the synthetic processes for the vitamins A, B, C, and E, solvent extraction is generally involved either in the separation steps for intermediates or in purification of the final

product, which is generally required in extremely high purity.

c. Miscellaneous pharmaceutical processes.

Solvent extraction is also used for preparation of many products that are either isolated from naturally occurring materials or purified during synthesis. Among these are sulfa drugs, methaqualone, phenobarbital, antihistamines, cortisone, estrogens and other hormones, and reserpine and other alkaloids. Distribution coefficient data for drug species are important for the design of solvent extraction procedures. A rapid determination of distribution coefficient data using a continuous solvent extraction system (AKUFVE) for pharmaceutical applications has been reported.

3. Food Processing

Solvent extraction is used in many processes in food processing. Industrial refining of fats and oils with propane is known as the Solxol process. Vegetable oils are refined by extraction using furfural as solvent. Solvent extraction is used in many protein manufacturing processes. Manufacture of fish protein by extracting the ground fish using isopropyl alcohol has been reported. Lactic acid for use as a food and drink additive is prepared from glucose solution by fermentation. The crude lactic acid is extracted from the fermentor broth using isopropyl ether as solvent.

4. Other Organic Processes

Solvent extraction has found application for many years in the coal tar industry. Extraction of phenols from coal-tar distillates by washing with caustic soda solution can be considered such a process. In the isomer separation, a process for separation of *m*- and *p*-cresol by dissociation extraction has been reported. Work is in progress in several parts of the world to use solvent extraction for the direct manufacture of chemicals from coal. Crude tall oil is a by-product of pulp mills. It is refined by solvent extraction using propane or furfural.

There are many applications where the organic compounds are extracted from natural materials. Extraction is used for preparation of pure flavor essences from expressed oils of various citrus fruit. Pyrethrum is recovered from pyrethrum flowers by solvent extraction. A continuous saponification, glycol extraction, and splitting process for converting fat into finished soap base has been used in soap production.

a. Industrial effluent treatment. Solvent extraction appears to have great potential in the field of effluent treatment, both for the economical recovery of valuable materials and for their removal to comply with statutory requirements. The Phenox process removes phenol from the effluent of catalytic cracking in petroleum refinery. Extraction processes may show a small profit from the value of the extracted phenols from ammoniacal coke-oven liquor. Oils are recovered by extraction from oily waste water from petroleum and petrochemical operations. Solvent extraction is employed commercially for the recovery of valuable by-products from the effluents produced in the wool industry and is applied in the same way in the pharmaceutical industry. Several solvent extraction schemes have been reported for organic industrial wastewater treatment.

b. Extractive reactions. Aromatic nitration in the manufacture of intermediates for dyestuffs, pharmaceutical products, plastics, and explosives is an example of an industrially important liquid-liquid organic reaction.

c. Supercritical-fluid extraction. Supercritical-fluid extraction (SCFE) has received great interest. In SCFE, the solvent is used at a temperature above the vapor-liquid critical point. It offers potential advantages: (a) enhanced transport properties—solute diffuses more rapidly through a supercritical solvent than through a liquid solvent; (b) equilibrium ratio and separation factors are generally quite high; and (c) the solvent can be recovered as a gas by reducing the pressure, which offers the prospect of significant energy saving.

Supercritical-fluid extraction has the advantage that slight changes in temperature and pressure within the critical region give extremely large changes in solvent density and solubility. There is greater flexibility in the process operating parameters of pressure and temperature as compared with conventional liquid-liquid extraction processes.

The disadvantage of SCFE is that the capital cost of a SCFE plant is substantially higher (at least 50%) than a conventional extraction plant.

Because of its low cost, nonhazardous chemical nature, and low critical temperature, carbon dioxide has been used in many applications. A commercial process to remove caffeine from coffee, using supercritical CO₂ as the solvent, is shown in Fig. 17. While actually a liquid-solid extraction process, it demonstrates principles involved in SCFE. A commercial SCFE process has been reported for recovery of hydrocarbon liquid from heavy oil. As compared with conventional propane deasphalting, this SCFE process can reduce capital and energy costs.

Potential new applications of SFE include the hydrometallurgical extraction of gold and the remediation of contaminated river sediments.

d. Biopolymer extraction. Rapid development of biotechnology has led to an intensification of efforts to

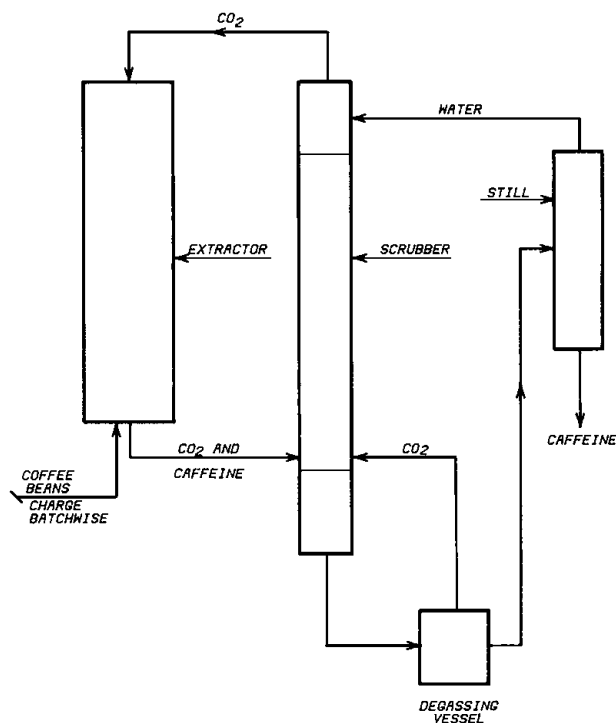


FIGURE 17 Supercritical-fluid extraction in decaffeination process.

develop new techniques for the separation of biochemicals from fermentation broth and cell culture media. While many methods exist for biochemical separation, most are limited to small-scale applications. Recent efforts have looked to solvent extraction as a technique which can be employed in a continuous mode and which can be applied on a large-scale production. Solvent extraction of low-molecular weight biochemicals such as antibiotics, using a conventional aqueous–organic solvent system, has been well documented. The extraction of proteins and bioparticles using biphasic aqueous polymer systems has received increasing interest. A hydrophilic liquid–liquid extraction system (an aqueous two-phase system) is formed when certain polymers such as polyethylene glycol (PEG) are added to an aqueous solution. A range of enzymes (Table III) show favorable partition properties in the system based on PEG/dextran solution, and, in many cases, rapid and effective removal of contaminants and undesirable products such as nucleic acids and polysaccharides is achieved.

e. Difficult separations. Difficult separations are frequently uneconomical because they involve high operating costs. Scheibel proposed process modifications of solvent extraction for separation of substances having separation factors as low as 0.95–1.05 to make it economically practical by reduction of the load on solvent recovery. The separations of *m*- and *p*-cresol, linoleic and abietic components of tall oil, and the production of heavy water have been achieved.

B. Inorganic Processes

Inorganic substances inherently tend to be insoluble in organic solvents. However, many metals can be separated

and concentrated from aqueous solutions of their salts by solvent extraction with a solution of an organic extractant in a carrier solvent (diluent).

The extractant must be able to react with the metal salt to form an organometallic compound or complex which is soluble in the diluent. The equilibrium constant should be favorable enough to permit a high metal loading of the organic phase. It should also be possible to reverse the extraction by a change of pH or temperature, thus permitting the stripping of the metal as a purified and/or more concentrated aqueous solution compared to the aqueous feed. Although the loading capacity of the organic phase is increased by raising the extractant concentration, it cannot be increased too much because of increased viscosity or more difficult phase separation.

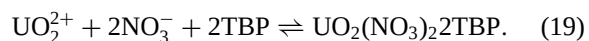
The diluent is chosen for its ease of phase separation after extraction, low cost, and safety in handling (low toxicity and flammability). Ordinarily the diluent plays no chemical role, although in some cases the metal extraction equilibria have been found to depend on whether an aromatic or an aliphatic diluent is used.

1. Nuclear Industry

Solvent extraction was first applied to metal separation in the nuclear industry in the late 1940s. Nuclear power generation by uranium fission produces “spent fuel” containing ^{238}U , ^{235}U , ^{239}Pu , ^{232}Th , and many other radioactive elements collectively known as fission products.

The uranium and plutonium are recovered for further use by first dissolving the spent fuel in nitric acid and subjecting the resulting solution to a solvent extraction process. Several different processes exist, the best known being the Purex process (Fig. 18), in which tributyl phosphate (TBP) (30% solution in kerosene) is the extractant. Extraction is carried out in compact mixer–settlers or air-pulsed columns fabricated of stainless steel, with about 99.9% removal of uranium and plutonium in the extract.

The governing equation for uranium extraction (solvating extraction) by TBP is



The loaded organic phase is subjected first to a stripping process under reducing conditions, thus removing the plutonium in its trivalent form. Then a second stripping operation is carried out with dilute nitric acid, which forms a complex with TBP, thereby driving the equilibrium of Eq. (19) to the left and transforming uranium back to the aqueous phase in pure form. The remaining TBP–kerosene mixture is washed and recycled for extraction of more dissolved spent fuel, as shown in Fig. 18. The design of nuclear fuel extraction plants is, of

TABLE III Enzymes with Favorable Partition Properties in Systems Based on PEG/Dextran/Salt Solutions

Enzyme	Applications
Alpha-amylase	Glues/food ingredients
Glucoamylase	Cornstarch/glucose conversion; starch/glucose conversion
Alpha-glucosidase	Maltose/glucose conversion
Glucose-6-phosphate dehydrogenase	Medicinal indicator
Formate dehydrogenase	Oxalate/formate determination
Formaldehyde dehydrogenase	Aldehyde/alcohol conversion
Catalase	Cold milk sterilization
Pullulanase	Starch/maltose conversion
Glucose isomerase	Glucose/fructose conversion
Beta-glucosidase	Food processing
Interferon	Pharmaceutical applications

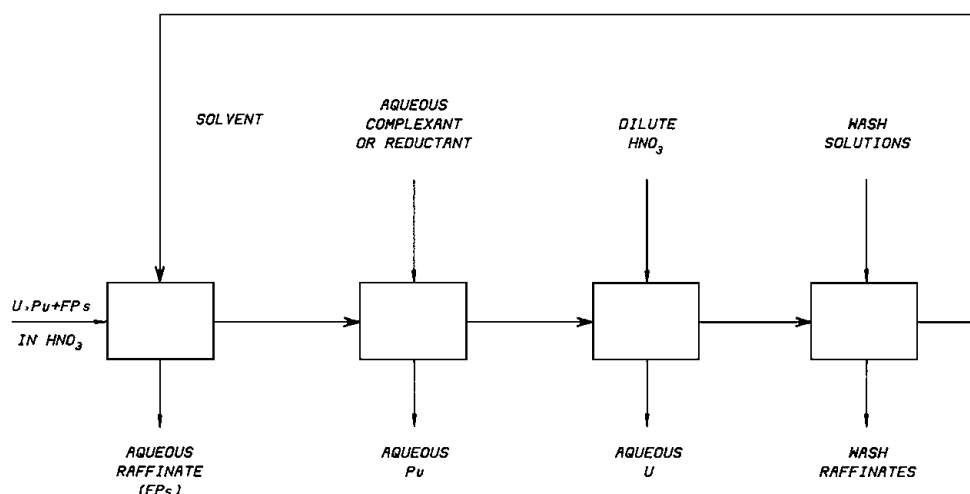


FIGURE 18 Solvent extraction process for spent nuclear fuel. [Reprinted with permission from Lo, T. C., Baird, M. H. I., and Hanson, C. (eds.) (1983). "Handbook of Solvent Extraction," Wiley (Interscience), New York. © 1983 John Wiley & Sons, Inc.]

course, subject to the same requirements as for all nuclear technology: containment, shielding, and the limitation of criticality.

Solvent extraction plays an important role in many commercial processes for the extraction of uranium from ore. In this case, the radioactivity levels are quite low compared with those in spent fuel extraction. The liquors from hydrometallurgical leaching of ores are typically fairly dilute in uranium (0.5–5 g/L) and contain iron and other metals in solution. Depending on conditions, solvent extraction or ion exchange may be used to separate and concentrate the uranium from the leach liquor.

2. Copper

The largest hydrometallurgical application of solvent extraction is in the refining of copper. Sulfuric acid leach liquors typically contain 1–5 g/L of copper and must be concentrated and purified to give a copper solution that can be subjected to electrowinning. Solvent extraction of copper from the acid leach liquors is carried out using a cation exchange reaction of the type shown in Eq. (5). The diluent is typically a low-cost kerosene. Many types of commercial copper extractants have been developed to meet the high demand. These may contain hydroxyoximes, carboxylic acids, or other compounds with an exchangeable hydrogen atom. There are about 20 major solvent extraction plants for copper in the world with capacities in the range 5000–80,000 ton/year. The extraction equilibria for copper are favorable enough that only a few mixer-settler stages are needed. A major element in the plant cost is settler size, which strongly affects the inventory of solvent required. Another cost factor is loss of solvent by entrain-

ment in the raffinate, which must be minimized by careful design of the mixers and settlers.

3. Nickel and Cobalt

Nickel and cobalt often occur with copper, and must be separated in pure form from hydrometallurgical leach liquors. Organic acid extractants can quite readily separate copper from cobalt and nickel, but the separation of cobalt from nickel is rather difficult. In one Ni/Co separation process, di-2-ethyl hexyl phosphoric acid (D2EHPA) is used as extractant, with strict control of the pH of the aqueous phase to take full advantage of the slightly different equilibrium constants for the Co and Ni reactions. Pulsed column contactors are used rather than mixer-settlers, and nickel impurity is removed from the loaded organic phase by scrubbing it with a cobalt-rich phase.

4. Other Metals

Solvent extraction is used in the separation of many other metals: some typical processes are listed in Table IV.

5. Phosphoric Acid

As well as the broad category of metal extraction processes, there are several inorganic processes involving the extraction of acids. Phosphoric acid has some solubility in organic solvents such as higher alcohols, ethers, and alkyl phosphates, and this can be utilized in several ways.

Crude phosphoric acid obtained from the wet process (action of dilute sulfuric acid on phosphate rock) contains many impurities such as fluorine, metals, and

TABLE IV Some Solvent Extraction Processes for Metals Other Than Uranium, Copper, Nickel, and Cobalt

Metal	Aqueous feed	Solvent used and notes on process
Tungsten	Sodium tungstate solution at pH 2	High-molecular weight amine extractant in kerosene; mixer-settler system
Molybdenum	Sodium molybdate solution at pH 4.5	High-molecular weight amine extractant in aromatic petroleum solvent; mixer-settler system
Vanadium	Raffinates from uranium ore extraction process Mineral leach liquors	D2EHPA and TBP in kerosene High-molecular weight amines in kerosene
Rare earths	Various mixed solutions	Different processes involving 2DEHPA, TBP, tertiary carboxylic acids, quaternary ammonium compounds, etc.
Thorium	Thorium nitrate solution with uranium and rare earths as impurities	TBP in kerosene followed by scrubbing and back extraction
Platinum group metals	Mixtures of anionic or neutral chlorocomplexes	Extractants: either anion exchangers (e.g., long-chain amino acids) or ligands (e.g., alkyl sulfides)
Gallium	Alkaline liquors from bauxite leaching (Bayer process)	Extractants; 8-hydroxyquinoline derivatives
Precious metals	Mixed chloride solution (Au, Pt, Pd)	8-Hydroxyquinoline derivatives in aromatic solvent
Arsenic	Acid leach liquors	Tributyl phosphate or higher alcohol extractant removes arsenic as impurity from other metals

organic matter. Purification of the phosphoric acid is carried out by extraction with a suitable alcohol or ether, followed by stripping of the organic phase at an elevated temperature.

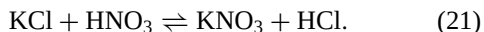
In some cases it may be economically desirable to obtain phosphoric acid from phosphate rock using hydrochloric acid instead of sulfuric acid,



The main problem here is the separation of phosphoric acid and calcium chloride, both of which are water soluble. This is carried out by solvent extraction of the phosphoric acid, followed by stripping (washing) with water. Some hydrochloric acid is also extracted in this process, and the phosphoric acid is finally purified and brought to 95% strength by evaporation.

6. Miscellaneous Inorganic Processes

Several other inorganic salt reactions with acids are promoted by solvent extraction and form the bases of processes or proposed processes. Prominent among these is the production of potassium nitrate from potassium chloride,



This reaction is favored by extracting hydrochloric acid continuously from the reacting aqueous phase, using a C₄ or C₅ alcohol solvent. The solvent also has an affinity for nitric acid, so the process includes a circuit for separate recovery of nitric acid and its return to the reaction stage.

IV. RECENT ADVANCES

This section describes some extraction techniques which are not yet in general use but show considerable potential.

A. Membrane Extraction

Conventional extraction from aqueous solutions requires a large inventory of organic solvent, which can be a major cost factor, particularly when an expensive extractant is needed as in the case of metal extraction. This has led to the development of membrane extraction techniques in which a relatively small amount of the organic phase is interposed between the aqueous feed and an aqueous strip solution. Figure 19 shows two techniques of membrane extraction. In emulsion membrane (ELM) extraction (Fig. 19a), the strip solution is present as emulsified drops in globules of the organic phase, which are contacted with the feed solution. This method requires a circuit in which the emulsion globules are broken up and separated so that the organic phase can be recycled. An alternative technique is supported membrane (SLM) extraction (Fig. 19b), in which the organic phase is supported on a porous solid wall. The use of a bundle of porous hollow fibers is effective in providing a large ratio of contact area to volume.

Membrane extraction can involve a chemical reaction, particularly in the extraction of metals [e.g., Eq. (5)]. In this case the extractant which is present in the membrane phase acts as a “carrier” taking the extracted species selectively across the membrane. This process is sometimes referred to as *pertraction*.

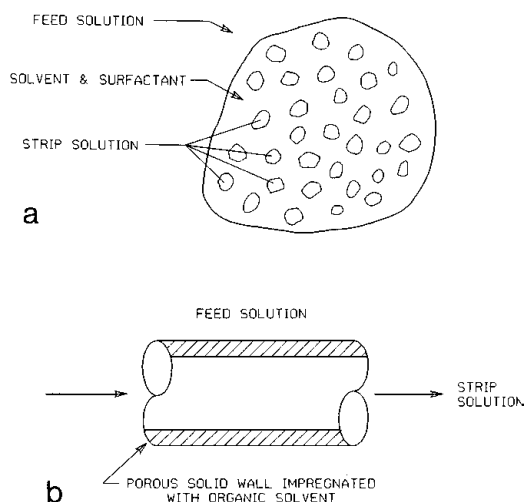


FIGURE 19 Liquid membrane extraction. (a) Emulsion membrane extraction, (b) supported membrane extraction.

B. Electrostatically Aided Extraction and Coalescence

Extraction can be enhanced by the application of a dc or pulsed electric field, typically on the order of 1 kV/cm. This requires that the aqueous phase be dispersed and the organic phase be of low conductivity. The improvement in mass transfer rate is due to the breakup of large drops by the action of the field and to the increase of drop velocity resulting in increased mass transfer coefficients. It has also been found that low-frequency pulsed fields are effective in breaking up emulsions in the settler stage of mixer-settler units.

C. Solvent-Impregnated Resins

As noted already, one of the greatest difficulties in working with liquid-liquid dispersions is the efficient separation of the phases after contact. An inefficient phase separation can lead to unacceptable losses of solvent and extractant. One way around this problem is to support the solvent phase on particles of an appropriate polymeric resin, which is then contacted with the feed phase. The phase separation (after contact) is facilitated and the losses of solvent and extractant are reduced. Initial studies were

made on the support of organic solvents/extractants on resin particles in a bulk aqueous phase, but resins are now available which can support an aqueous extractant in a bulk organic phase.

SEE ALSO THE FOLLOWING ARTICLES

CHEMICAL THERMODYNAMICS • COORDINATION COMPOUNDS • ELECTROLYTE SOLUTIONS, TRANSPORT PROPERTIES • FLUID MIXING • MINERAL PROCESSING • PETROLEUM REFINING • PHARMACEUTICALS • RUBBER, SYNTHETIC • SYNTHETIC FUELS

BIBLIOGRAPHY

- Godfrey, J. C., and Slater, M. J. (eds.). (1994). "Liquid-Liquid Extraction Equipment," Wiley, Chichester, UK.
- "International Solvent Extraction Conference, ISEC '86, Munich, Reprints," Vols. 1-3, Dechema, Frankfurt (1986).
- "International Solvent Extraction Conference, ISEC '88, Moscow, Proceedings," Vols. 1-4, USSR Academy of Sciences (1988).
- "International Solvent Extraction Conference, ISEC '90, Kyoto, Proceedings," Elsevier.
- "International Solvent Extraction Conference ISEC '93, York, Proceedings," Vols. 1-3, Elsevier, Amsterdam (1993).
- "International Solvent Extraction Conference ISEC '96, Melbourne, Proceedings," Vols. 1, 2, University of Melbourne.
- "International Solvent Extraction Conference ISEC '99, Barcelona, Book of Abstracts" (1999).
- Lo, T. C. (1997). In "Handbook of Separation Techniques for Chemical Engineers," 3rd ed. (P. Schweitzer, ed.), McGraw-Hill, New York.
- Lo, T. C., and Baird, M. H. I. (1994). In "Encyclopedia of Chemical Technology," 4th ed., Vol. 10, pp. 125-180, Wiley (Interscience), New York.
- Lo, T. C., Baird, M. H. I., and Hanson, C. (eds.). (1983). "Handbook of Solvent Extraction," Wiley (Interscience), New York.
- Ritcey, G. M., and Ashbrook, A. W. (1978). "Solvent Extraction—Principles and Applications to Process Metallurgy," Vol. 1, Elsevier, Amsterdam.
- Ritcey, G. M., and Ashbrook, A. W. (1984). "Solvent Extraction—Principles and Applications to Process Metallurgy," Vol. 2, Elsevier, Amsterdam.
- Rydberg, J., Musikas, C., and Choppin, G. R. (eds.). (1992). "Principles and Practices of Solvent Extraction," Marcel Dekker, New York.
- Thornton, J. D. (ed.). (1992). "Science and Practice of Solvent Extraction," Oxford University Press, Oxford.
- Wisniak, J., and Tamir, A. (1980). "Liquid-Liquid Equilibrium and Extraction," Elsevier, Amsterdam.



Surfactants, Industrial Applications

Tharwat F. Tadros

Imperial Chemical Industries

- I. Introduction
- II. General Classification of Surfactants
- III. Physical Properties of Surfactant Solutions
- IV. Adsorption of Surfactants at Various Interfaces
- V. Surfactants as Emulsifiers
- VI. Surfactants as Dispersants
- VII. Role of Surfactants in Stabilization of Emulsions and Suspensions
- VIII. Role of Surfactants in Solubilization and Microemulsions

- IX. Surfactants in Foams
- X. Surfactants in Wetting Phenomena
- XI. Application of Surfactants In Cosmetics and Personal Care Products
- XII. Application of Surfactants in Pharmaceuticals
- XIII. Application of Surfactants in Agrochemicals
- XIV. Application of Surfactants in the Food Industry

GLOSSARY

Cloud point Temperature at which a surfactant solution of a given concentration becomes turbid (cloudy).

Critical micelle concentration (cmc) Concentration at which the physical properties of a surfactant solution show an abrupt change.

Emulsion Dispersion of a liquid in a liquid.

Krafft temperature Temperature at which the surfactant shows a sudden increase in solubility.

Micelles Association units of surfactant molecules.

Microemulsion Thermodynamically isotropic system of water, oil, and surfactant.

Solubilization Incorporation of an insoluble substance in a surfactant solution.

Suspension Dispersion of a solid in a liquid.

SURFACE-ACTIVE agents (usually referred to as surfactants) are amphipathic molecules consisting of a non-polar hydrophobic portion, usually a straight or branched hydrocarbon or fluorocarbon chain containing 8–18 carbon atoms, which is attached to a polar or ionic portion (hydrophilic). The hydrophilic portion can, therefore, be nonionic, ionic, or zwitterionic, accompanied by counterions in the last two cases. The hydrocarbon chain interacts weakly with the water molecules in an aqueous environment, whereas the polar or ionic head group interacts strongly with water molecules via dipole or ion–dipole interactions. It is this strong interaction with the water molecules which renders the surfactant soluble in water. However, the cooperative action of dispersion and hydrogen bonding between the water molecules tends to “squeeze” the hydrocarbon chain out of the water and

hence these chains are referred to as hydrophobic. The balance between hydrophobic and hydrophilic parts of the molecule gives these systems their special properties, e.g., accumulation at various interfaces and association in solution.

Surfactants find application in almost every chemical industry, of which the following are worth mentioning: detergents, paints, dyestuffs, personal care and cosmetics, pharmaceuticals, agrochemicals, ceramics, fibres, plastics, and paper coating. Moreover, surfactants play a major role in the oil industry, for example, in enhanced and tertiary oil recovery. They are also occasionally used for environmental protection, e.g., in oil slick dispersants. Therefore, a fundamental understanding of the physical chemistry of surface active agents, their unusual properties, and their phase behavior is essential for most industrial chemists. In addition, understanding the basic phenomena involved in the application of surfactants such as in the preparation of emulsions and suspensions and their subsequent stabilization, in microemulsions, and in wetting, spreading, and adhesion, is of vital importance in arriving at the right composition of many industrial formulations.

I. INTRODUCTION

In this overview, I start with the general classification of surfactants and their unusual properties. This is followed by some examples to illustrate the application of surfactants in some chemical industries.

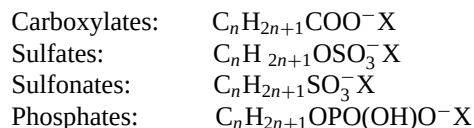
II. GENERAL CLASSIFICATION OF SURFACTANTS

A simple classification of surfactants based on the nature of the hydrophilic group is commonly used. Four main classes may be distinguished, namely, anionic, cationic, zwitterionic, and nonionic. A useful technical reference is McCutchen. Another useful text, by van Oss *et al.*, gives a list of the physicochemical properties of selected anionic, cationic, and nonionic surfactants. The handbook by Porter is also a useful book for classification of surfactants. Another important class of surfactants, which has attracted considerable attention in recent years, is the polymeric type. A brief description of the various classes is given below.

A. Anionic Surfactants

This is the most widely used class of surfactants in industrial application, due to their relatively low cost of manu-

facture and their wide application in detergents. For optimum detergency, the hydrophobic chain is a linear alkyl chain with a length in the region of 12–16 C atoms. Linear chains are preferred since they are more effective and more degradable than branched ones. The most commonly used hydrophilic groups are carboxylates, sulfates, sulfonates and phosphates. A general formula may be ascribed to anionic surfactants as follows.



n is usually in the range 8–18 C atoms, and the counterion X is usually Na^+ .

Several other anionic surfactants are commercially available, such as sulfosuccinates, isethionates, and taurates, and these are sometimes used for special applications. The carboxylates and sulfates are sometimes modified by the incorporation of a few moles of ethylene oxide (referred to as ether carboxylates and ether sulfates, respectively).

B. Cationic Surfactants

The most common cationic surfactants are the quaternary ammonium compounds with the general formula $R'R''R'''R''''NX^+$, where X is usually chloride ion and R represents alkyl groups. A common class of cationics is the alkyl trimethyl ammonium chloride, where R contains 8–18 C atoms, e.g., dodecyl trimethyl ammonium chloride, $C_{12}H_{25}(CH_3)_3NCl$. Another widely used cationic surfactant class is that containing two long-chain alkyl groups, having a chain length of 8–18 C atoms. These dialkyl surfactants are less soluble in water and they produce multilamellar structures (vesicles or liposomes). They are commonly used as fabric softeners.

A widely used cationic surfactant is benzalkonium chloride, where one of the methyl groups is replaced by a benzyl group, $C_6H_5-CH_2$. This surfactant is commonly used as a bactericide. Cationic surfactants can also be modified by the incorporation of ethylene oxide chains in the molecules.

C. Amphoteric Surfactants

These are surfactants containing both cationic and anionic groups. The most common amphoteric are the N -alkyl betaines, which are derivatives of trimethylglycine, $(CH_3)_3NCH_2COOH$ (which was described as betaine). An example of a betaine surfactant is laurylamidopropyl dimethyl betaine, $C_{12}H_{25}CON(CH_3)_2CH_2COOH$.

The main characteristics of amphoteric surfactants is their dependence on the pH of the solution in which they are dissolved. In acid solutions, the molecule acquires a positive charge and it behaves like a cationic, whereas in alkaline solutions they become negatively charged and behave like an anionic. A specific pH can be defined at which both ionic groups show equal ionization (the isoelectric point of the molecule).

D. Nonionic Surfactants

The most common nonionic surfactants are those based on ethylene oxide, referred to as ethoxylated surfactants. Several classes can be distinguished: alcohol ethoxylates, alkyl phenol ethoxylates, fatty acid ethoxylates, sorbitan ester ethoxylates, fatty amine ethoxylates, and ethylene oxide-propylene oxide copolymers (sometimes referred to as polymer surfactants). Another important class of nonionics are the multihydroxy products such as glycol esters, glycerol (and polyglycerol) esters, glucosides (and polyglucosides), and sucrose esters. Amine oxides and sulfanyl surfactants represent nonionic with a small head group.

The critical micelle concentration (cmc) of nonionics (see below) is about two orders of magnitude lower than the corresponding anionics with the same alkyl chain length. Molecules with an average alkyl chain length of 12 C atoms and containing more than 5 ethylene oxide (EO) units are usually soluble in water at room temperature. However, as the temperature of the solution is gradually increased, the solution becomes cloudy (as a result of dehydration of the PEO chain) and the temperature at which this occurs is referred to as the cloud point (CP) of the surfactant solution. At any given alkyl chain length the CP increases with increase in the number of EO units in the molecule. The CP is also affected by the presence of electrolytes. Generally, the CP decreases with an increase in electrolyte concentration. The reduction in cloud point depends also on the nature of the electrolyte added.

E. Fluorocarbon and Silicone Surfactants

These surfactants can lower the surface tension of water, γ , to values below 20 mN m^{-1} (most surfactants lower γ to values in the region of 30 mN m^{-1}) and hence they are sometimes described as superwetters. They are very useful for enhancing the wetting and spreading of liquids on solid substrates.

F. Polymeric Surfactants

Polymeric surfactants are specially designed to produce excellent dispersing agent (for suspensions) and emulsi-

fiers (for emulsions). Several molecules have been introduced such as the block copolymers of polyethylene oxide (PEO) and polypropylene oxide (PPO). These are A–B–A block copolymers with the structure PEO–PPO–PEO and molecules are commercially available with various proportions of PEO and PPO. Other types of A–B block copolymers are those based on polystyrene (PS) and PEO. Graft copolymers with one B chain and several A chains also exist such as a graft of poly(methylmethacrylate) with a number of PEO side chains. These polymeric surfactants have been applied for the preparation of highly concentrated suspensions and emulsions to enhance their long term stability.

G. The Hydrophilic–Lipophil Balance (HLB)

A useful index for choosing surfactants for various applications is the hydrophilic–lipophilic balance (HLB), which is based on the relative percentage of hydrophilic-to-lipophilic groups in the surfactant molecule(s). Surfactants with a low HLB number normally form W/O emulsions, whereas those with a high HLB number form a O/W emulsion. A summary of the HLB range required for various purposes is given in Table I.

Griffin developed a simple equation for calculation of the HLB number of certain numbers of nonionic surfactants such as fatty acid esters and alcohol ethoxylates. For the polyhydroxy fatty the HLB number is given by the equation

$$\text{HLB} = 20 \left(1 - \frac{S}{A} \right), \quad (1)$$

where S is the saponification number of the ester and A is the acid number of the acid. Thus, a glycerol monostearate, with $S = 161$ and $A = 198$, will have an HLB number of 3.8, i.e., it is suitable for a W/O emulsifier.

For the simpler alcohol ethoxylates, the HLB number can be calculated from the weight percentages of oxyethylene E and polyhydric alcohol P , i.e.,

$$\text{HLB} = \frac{(E + P)}{5}. \quad (2)$$

TABLE I Summary of HLB Ranges for Various Applications

HLB range	Application
3–6	W/O emulsifier
7–9	Wetting agent
8–18	O/W emulsifier
13–16	Detergent
15–18	Solubilizer

III. PHYSICAL PROPERTIES OF SURFACTANT SOLUTIONS

The physical properties of surface active agents differ from those of smaller or nonamphipathic molecules in one major aspect, namely, the abrupt changes in their properties above a critical concentration. This is illustrated in Fig. 1, in which a number of physical properties (surface tension, osmotic pressure, turbidity, solubilization, magnetic resonance, conductivity, and self-diffusion) are plotted as a function of concentration. All these properties (interfacial and bulk) show an abrupt change at a particular concentration, which is consistent with the fact that above this concentration, surface active ions or molecules in solution associate to form larger units. These association units are called micelles and the concentration at which this association phenomenon occurs is known as the critical micelle concentration (cmc).

Each surfactant molecule has a characteristic cmc value at a given temperature and electrolyte concentration. A compilation of cmc values was given in 1971 by Mukerjee and Mysels. As an illustration, the cmc values of a number of surfactants are given in Table II, to show some of the general trends. Within any class of surface active agents, the cmc decreases with an increase in chain length of the hydrophobic chain. With nonionic surfactants, increasing the length of the hydrophilic (PEO) chain causes an increase in the cmc. In general, nonionic surfactants have lower cmc values than their corresponding ionic surfactants with the same chain length. Incorporation of a phenyl group in the alkyl chain increases its hydrophobicity to a much smaller extent than increasing its chain length with the same number of C atoms.

The presence of micelles can account for many of the unusual properties of solutions of surfactants. For example, it can account for the near-constant surface tension

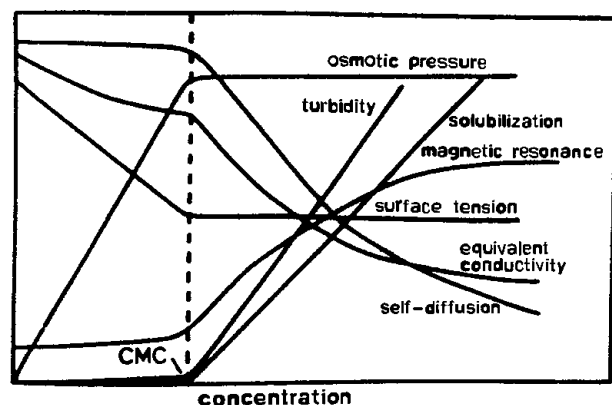


FIGURE 1 Changes in the concentration dependence of a wide range of physicochemical changes around the cmc. [From Lindman, B. (1984). *Surfactants*. Academic Press, London.]

TABLE II The cmc Values of Some Surfactants

Surface-active agent	cmc (mol dm ⁻³)
(A) Anionic	
Sodium octyl-1-sulfate	1.30×10^{-1}
Sodium decyl-1-sulfate	3.32×10^{-2}
Sodium dodecyl-1-sulfate	8.39×10^{-3}
Sodium tetradecyl-1-sulfate	2.05×10^{-4}
(B) Cationic	
Octyl trimethyl ammonium bromide	1.30×10^{-1}
Decyl trimethyl ammonium bromide	6.46×10^{-2}
Dodecyl trimethyl ammonium bromide	1.56×10^{-2}
Hexadecyl trimethyl ammonium bromide	9.20×10^{-4}
(C) Nonionic	
Octyl hexaoxyethylene glycol monoether, C ₈ E ₆	9.80×10^{-3}
Decyl hexaoxyethylene glycol monoether, C ₁₀ E ₆	9.00×10^{-4}
Decyl nonaoxyethylene glycol monoether, C ₁₀ E ₉	1.30×10^{-3}
Dodecyl hexaoxyethylene glycol monoether, C ₁₂ E ₆	8.70×10^{-5}
Octylphenyl hexaoxyethylene glycol monoether, C ₈ φE ₆	2.05×10^{-4}

value above the cmc. It can also account for the reduction in molar conductance above the cmc, the rapid increase in turbidity above the cmc, etc.

The size and shape of micelles have been a subject of several debates. It is now generally accepted that three main shapes of micelles are present, depending on the surfactant structure and the environment in which they are dissolved, e.g., electrolyte concentration and type, pH, and presence of nonelectrolytes. The most common shape of micelles is a sphere with the following properties: (i) an association unit with a radius approximately equal to the length of the hydrocarbon chain (for ionic micelles); (ii) an aggregation number of 50–100 surfactant monomers; (iii) bound counterions for ionic surfactants; (iv) a narrow range of concentrations at which micellization occurs; and (v) a liquid interior of the micelle core.

Two other shapes of micelles may be considered, namely, the rod-shaped micelle suggested by Debye and Anacker and the lamellar micelle suggested by McBain. The rod-shaped micelle was suggested to account for the light-scattering results of cetyl trimethyl ammonium bromide in KBr solutions, whereas the lamellar micelle was considered to account for the X-ray results in soap solutions. A schematic picture of the three type of micelles is shown in Fig. 2.

One of the characteristic features of solutions of surfactants is their solubility–temperature relationship, which is illustrated in Fig. 3 for an anionic surfactant, namely, sodium decyl sulfonate. It can be seen that the solubility of the surfactant increases gradually with an increase in temperature, but above 22°C there is a rapid increase in

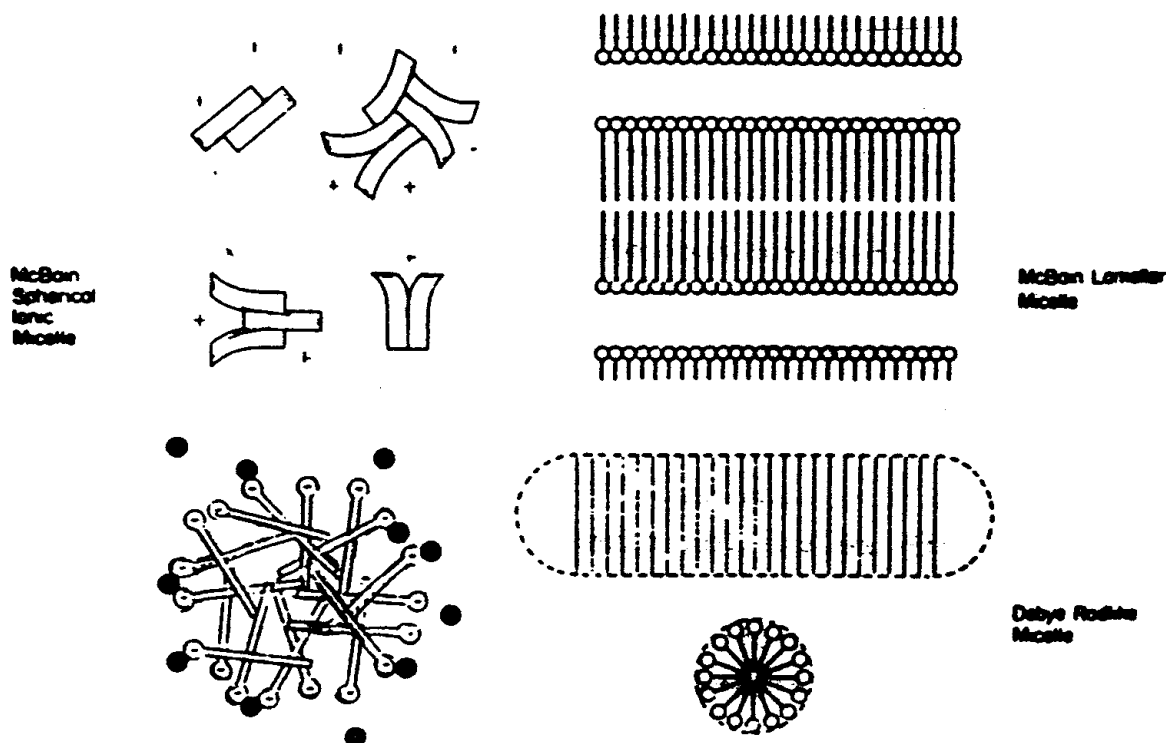


FIGURE 2 Various shapes of micelles following McBain (II). [Adapted from Hartley (1936) and Debye and Anaker (1951).]

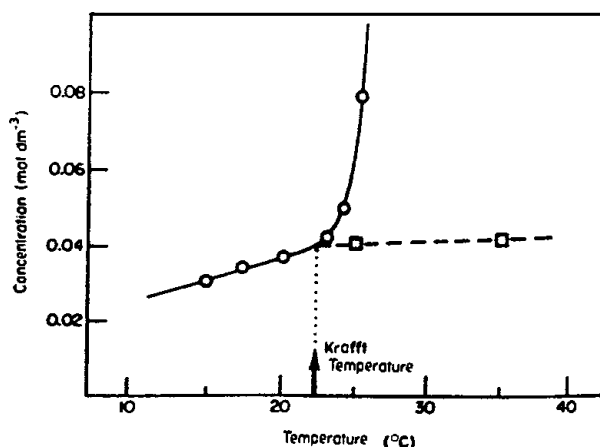


FIGURE 3 Solubility (○) and cmc (□) versus temperature for the sodium decyl sulfonate–water system.

solubility with further increases in temperature. The cmc of the surfactant also increases slowly with increases in temperature. The point at which the solubility curve intersects with the cmc curve (i.e., solubility = cmc) is referred to as the Krafft temperature of the surfactant. At the Krafft temperature, there is an equilibrium among solid hydrated surfactant, micelles, and monomer. The Krafft temperature of an ionic surfactant increases with an increase in the alkyl-chain length. For that reason, surfactants with

an alkyl chain longer than 12–14 C atoms are generally not very useful for application, since a concentrated solution can be prepared only at temperatures significantly higher than room temperature. One useful way to reduce the Krafft temperature is to use an alkyl chain with a wide chain length distribution. Indeed most commercial surfactants have this wide range since they are produced from natural fats and oils.

The solubility–temperature relationship for nonionic surfactants is different from that of ionic surfactants. This is illustrated in Fig. 4 which, shows the phase diagram for the binary system, dodecyl hexaoxyethylene glycol monoether ($C_{12}E_6$)–water. This phase diagram shows the various phases that are formed when the surfactant concentration and temperature are changed. The cloud-point curve at the top of the phase diagram separates the 2L (two liquid phases appear above the cloud point, one rich in surfactant and the other rich in water) from the I isotropic solution. Below the cloud-point curve, the phase diagram shows some characteristic regions at high surfactant concentrations, namely, the M and N region. The M-phase is the region of hexagonal or middle phase, which consists of cylindrical units that are hexagonally close-packed. In this region, the viscosity of the surfactant solution is extremely high and the system appears like a transparent gel. However, when viewed under the polarizing microscope,

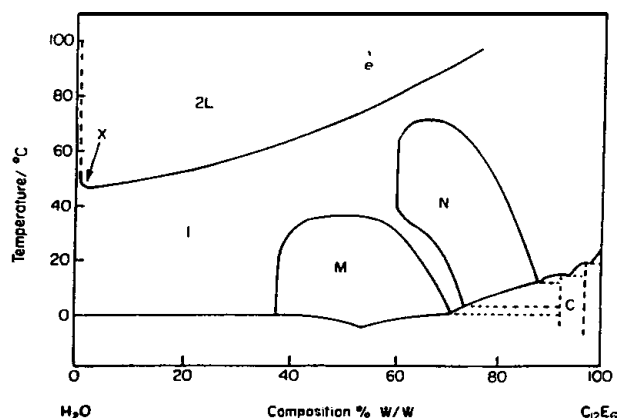


FIGURE 4 Phase diagram for the dodecyl hexaoxyethylene glycol monoether–water system.

it shows some texture (“fan-like structure”), which is due to the anisotropy of the units formed. The N-phase is the lamellar or neat phase, which consists of sheets of molecules in a bimolecular packing with head groups exposed to the water layers between them. This phase is less viscous than the middle phase and it shows different textures under the polarizing microscope (“oily streaks” and “maltese crosses”). Several other liquid crystalline phases may be identified with other nonionic surfactants, such as the cubic viscous isotropic phase (which shows no texture under the polarizing microscope). Figure 5 shows schematic pictures of the three phases described above.

A. Thermodynamics of Micellization

Micellization is a dynamic phenomenon in which n monomeric surfactant molecules S associate to form a micelle S_n ,



Hartley envisaged a dynamic equilibrium whereby surface-active agent molecules are constantly leaving the

micelles while other molecules enter the micelle. Experimental techniques using fast kinetic methods such as stop flow, temperature and pressure jumps, and ultrasonic relaxation have shown that there are two relaxation processes for micellar equilibrium. The first relaxation time τ_1 is of the order of 10^{-7} sec (10^{-8} – 10^{-3} sec) and represents the lifetime of a surface active molecule in the micelle, i.e., it represents the association and dissociation rate for a single molecule entering and leaving the micelle. The second relaxation time τ_2 corresponds to a relatively slow process, namely, the micellization–dissolution process represented by Eq. (3). The value of τ_2 is of the order of milliseconds (10^{-3} –1 sec).

The equilibrium aspect of micelle formation can be considered by application of the second law of thermodynamics. The equilibrium constant for the process represented by Eq. (3) is given by

$$K = \frac{[S_n]}{S^n} = \frac{C_m}{C_s^n}, \quad (4)$$

where C_s and C_m represent the concentration of monomer and micelle respectively.

The standard free energy of micellization, ΔG^0 , is then given by

$$-\Delta G_m^0 = RT \ln K = RT \ln C_m - nRT \ln C_s, \quad (5)$$

and the free energy per monomer, $\Delta G^0 (= \Delta G_m^0/n)$, is given by

$$-\Delta G^0 = \left(\frac{RT}{n} \right) \ln C_m - RT \ln C_s. \quad (6)$$

For many micellar systems, n is a large number (>50), and therefore, the first term on the right-hand side of Eq. (6) may be neglected:

$$\Delta G^0 = RT \ln C_s = RT \ln [\text{cmc}]. \quad (7)$$

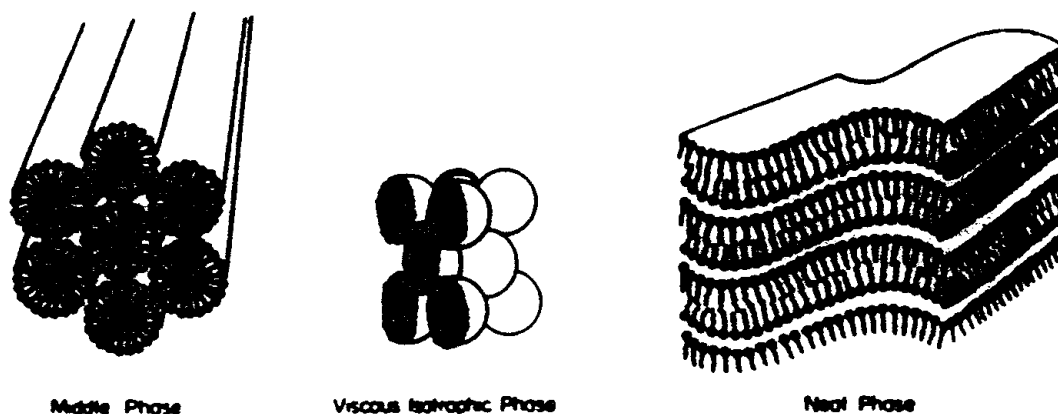


FIGURE 5 Schematic representation of the structures found in concentrated surfactant solutions.

ΔG^0 is always negative and this shows that micelle formation is a spontaneous process. For example, for $C_{12}E_6$, the cmc is $8.70 \times 10^{-5} \text{ mol dm}^{-3}$ and $\Delta G^0 = -33.1 \text{ KJ mol}^{-1}$ (expressing the cmc as the mole fraction).

The enthalpy of micellization ΔH^0 can be measured either from the variation of cmc with temperature or directly by microcalorimetry. From ΔG^0 and ΔH^0 , one can obtain the entropy of micellization ΔS^0 ,

$$\Delta G^0 = \Delta H^0 - T \Delta S^0. \quad (8)$$

Measurement of ΔH^0 and ΔS^0 showed that the former is small and positive and the second is large and positive. This implies that micelle formation is entropy driven and is described in terms of the hydrophobic effect (14). Then hydrophobic chains of the surfactant monomers tend to reduce their contact with water, whereby the latter form “icebergs” by hydrogen bonding. This results in reduction of the entropy of the whole system. However, when the monomers associate to form micelles, these “icebergs” tend to melt (hydrogen bonds are broken), and this results in an increase in the entropy of the whole system.

IV. ADSORPTION OF SURFACTANTS AT VARIOUS INTERFACES

The adsorption of surfactants at the liquid/air interface, which results in surface tension reduction, is important for many applications in industry such as wetting, spraying, impaction, and adhesion of droplets. Adsorption at the liquid/liquid interface is important in emulsification and subsequent stabilization of the emulsion. Adsorption at the solid/liquid interface is important in wetting phenomena, preparation of solid/liquid dispersions, and stabilization of suspensions. Below a brief description of the various adsorption phenomena is given.

A. Adsorption at Air/Liquid and Liquid/Liquid Interfaces

Gibbs derived a thermodynamic relationship between the surface or interfacial tension γ and the amount of surfactant adsorbed per unit area at the A/L or L/L interface, Γ (referred to as the surface excess),

$$\frac{d\gamma}{d \ln C} = -\Gamma RT, \quad (9)$$

where C is the surfactant concentration (mol dm^{-3}).

Equation (9) allows one to obtain the surface excess from the variation of surface or interfacial tension with surfactant concentration. Γ can be obtained from the slope of the linear portion of the $\gamma - \log C$ curve as illustrated in Fig. 6 for A/L and L/L interfaces.

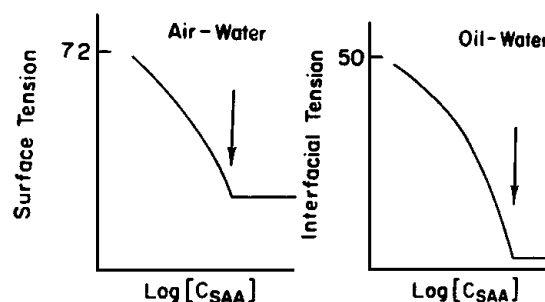


FIGURE 6 Variation of surface and interfacial tension with $\log [C_{SAA}]$ at the air–water and oil–water interface.

It can be seen that for the A/W interface γ decreases from the value for water ($\sim 72 \text{ mN m}^{-1}$), reaching about $25\text{--}30 \text{ mN m}^{-1}$ near the cmc. For the O/W interface γ decreases from $\sim 50 \text{ mN m}^{-1}$ (for a pure hydrocarbon–water interface) to $\sim 1\text{--}5 \text{ mN m}^{-1}$. Clearly the rate of reduction of γ with $\log C$ below the cmc and the limiting γ reached at and above the cmc depend on the nature of surfactant and the interface.

From Γ , the area per surfactant ion or molecule can be calculated:

$$\text{area/molecule } (A) = \frac{1}{\Gamma N_{av}} (\text{m}^2) = \frac{10^{18}}{\Gamma N_{av}} (\text{nm}^2). \quad (10)$$

The area/molecule A gives information on the surfactant orientation at the interface. For example, for an anionic surfactant such as sodium dodecyl sulfate, A is determined by the area occupied by the alkyl chain and head group, if these molecules lie “flat” at the interface. For a vertical orientation, A is determined by the area of the head group ($-\text{O}-\text{SO}_3^-$), which, at a low electrolyte concentration, is in the region of 0.4 nm^2 . This area is larger than the geometrical area occupied by a sulfate group, as a result of the lateral repulsion between the head groups. On the addition of electrolyte, this lateral repulsion is reduced and A reaches a smaller value ($\sim 0.2 \text{ nm}^2$). For nonionic surfactants, A is determined by the area occupied by the polyethylene oxide chain and A increases with an increase in the number of EO units (values in the region of $1\text{--}2 \text{ nm}^2$ are common with ethoxylated surfactants).

An important point can be made from the $\gamma - \log C$ curve. At a concentration just below the cmc, the curve is linear, indicating that saturation adsorption is reached just below the cmc. Above the cmc, the slope of the $\gamma - \log C$ curve is nearly zero, indicating a near-constant activity of the surfactant ions or molecules just above the cmc.

B. Adsorption of Surfactant at the Solid/Liquid Interface

The adsorption of surfactants at the S/L interface involves a number of complex interactions, such as hydrophobic,

polar, and hydrogen bonding. This depends on the nature of the substrate as well as that of the surfactant ions or molecules. Generally speaking, solid substrates may be subdivided into hydrophobic (nonpolar) and hydrophilic (polar) surfaces. The surfactants can be ionic or nonionic, and they interact with the surface in a specific manner. The adsorption of ionic surfactants on hydrophobic surfaces (such as C black, polystyrene, and polyethylene) is determined by hydrophobic bonding between the alkyl chain and the nonpolar surface. In this case, the charged or polar head groups play a relatively smaller role, except in their lateral repulsion, which reduces adsorption. For this reason, the addition of electrolyte to ionic surfactants generally results in an increase in adsorption. The same applies for nonionic surfactants, which also show an increase in adsorption with increasing temperature.

The adsorption of surfactants on solid substrates may be described by the Frumkin–Fowler–Guggenheim equation,

$$\frac{\theta}{(1-\theta)} \exp(A\theta) = \frac{C}{55.5} \exp\left(\frac{-\Delta G_{\text{ads}}^0}{kT}\right), \quad (11)$$

where θ is the fractional surface coverage, which is given by Γ/N_s (where Γ is the number of moles adsorbed per unit area and N_s is the total number of adsorption sites as moles per unit area for monolayer saturation adsorption), C is the bulk solution concentration as moles ($C/55.5$ gives the mole fraction of surfactant), A is a constant that is introduced to account for lateral interaction between the surfactant ions or molecules, and ΔG_{ads}^0 is the standard free energy of adsorption, which may be considered to consist of two contributions, an electrical term ΔG_{elec}^0 and a specific adsorption term ΔG_{spec}^0 . The latter may consist of various contributions arising from chain–chain interaction, ΔG_{cc}^0 , chain–surface interaction, ΔG_{cs}^0 , and head group–surface interaction, ΔG_{hs}^0 .

In many cases, the adsorption of surfactants on hydrophobic surfaces may follow a Langmuir-type isotherm,

$$\Gamma_2 = \frac{\Delta C}{mA} = \frac{abC_2}{1 + bC_2}, \quad (12)$$

where ΔC is the number of moles of surfactant adsorbed by m grams of adsorbent with surface area A ($\text{m}^2 \text{g}^{-1}$), C_2 is the equilibrium concentration, a is the saturation adsorption, and b is a constant related to the free energy of adsorption ($b \propto -\Delta G_{\text{ads}}^0$). The saturation adsorption a can be used to obtain the area per molecule A , as discussed above ($A = 1/a N_{\text{av}} \text{m}^2$ or $10^{18}/a N_{\text{av}} \text{nm}^2$).

The adsorption of ionic or polar surfactants on charged or polar surfaces involves coulombic (ion–surface charge interaction), ion–dipole, and/or dipole–dipole interaction. For example, a negatively charged silica surface (at a pH above the isoelectric point of the surface, i.e., pH >2–3)

will adsorb a cationic surfactant by interaction between the negatively charged silanol groups and the positively charged surfactant ion. The adsorption will continue till all negative charges on silica are neutralized and the surface will have a net zero charge (the surface becomes hydrophobic). When the surfactant concentration is further increased, another surfactant layer may build up by hydrophobic interaction between the alkyl chain of the surfactant ions, and the surface now acquires a positive charge and it becomes hydrophilic. However, the adsorption of ionic surfactants on hydrophilic surfaces may acquire additional features, whereby the surfactant ions may associate on the surface, forming “hemimicelles.” An example of this behavior is the adsorption of sodium dodecyl sulfonate (an anionic surfactant) on a positively charged alumina surface (at a pH below its isoelectric point, i.e., pH 7). Initially, the adsorption occurs by a simple ion-exchange mechanism whereby the surfactant anions exchange with the chloride counterions. In this region, the adsorption shows a slow increase with an increase in surfactant concentration. However, above a certain surfactant concentration (that is, just above that for complete ion exchange), the adsorption increases very rapidly with further increases in surfactant concentration. This is the region of hemimicelle formation, whereby several surfactant ions associate to form aggregation units on the surface.

The adsorption of nonionic surfactants on polar and nonpolar surfaces also exhibits various features, depending on the nature of the surfactant and the substrate. Three types of isotherms may be distinguished, as illustrated in Fig. 7. These isotherms can be accounted for by the different surfactant orientations and their association at the solid/liquid interface as illustrated in Fig. 8. Again, bilayers, hemimicelles, and micelles can be identified on various substrates.

V. SURFACTANTS AS EMULSIFIERS

Emulsions are a class of disperse systems consisting of two immiscible liquids, one constituting the droplets (the disperse phase) and the second the dispersion medium. The most common class of emulsions is those whereby the droplets constitute the oil phase and the medium is an aqueous solution (referred to as O/W emulsions) or where the droplets constitute the disperse phase, with the oil being the continuous phase (W/O emulsions). To disperse a liquid into another immiscible liquid requires a third component, referred to as the emulsifier, which in most cases is a surfactant. Several types of emulsifiers may be used to prepare the system, ranging from anionic, cationic, zwitterionic, and nonionic surfactants to more specialized emulsifiers of the polymeric type, referred to as polymeric

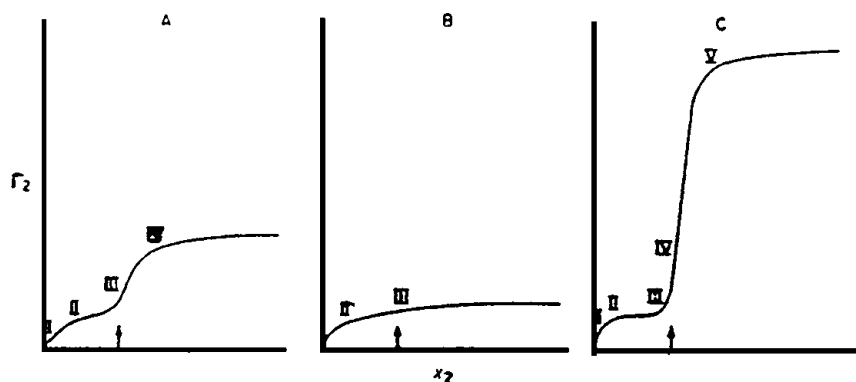


FIGURE 7 Adsorption isotherms corresponding to the three adsorption sequences shown in Fig. 8 (the cmc is indicated by the arrow).

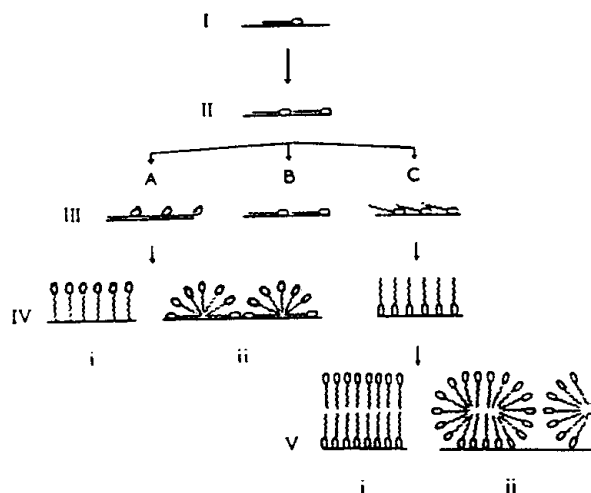


FIGURE 8 Model for the adsorption of nonionic surfactants showing the orientation of the molecules at the surface.

surfactants (see above). As discussed before, W/O emulsions require a low-HLB number surfactant, whereas for O/W emulsions a high-HLB number surfactant (8–18) is required.

The emulsifier plays a number of roles in the formation of the emulsion and subsequent stabilization. First, it reduces the O/W interfacial tension, thus promoting the formation of smaller droplets. More important is the result of the interfacial tension gradient $d\gamma/dz$, which stabilizes the liquid film between the droplets, thus preventing film collapse during emulsification. Another important role for the emulsifier is to reduce coalescence during emulsification as the result of the Gibbs–Marangoni effect. As a result of the incomplete adsorption of the surfactant molecules, an interfacial tension gradient $d\gamma/dA$ is present, and this results in a Gibbs elasticity, ε_f ,

$$\varepsilon_f = \frac{2\gamma(d \ln \Gamma)}{1 + \left(\frac{1}{2}\right) h(dC/d\Gamma)}, \quad (13)$$

where h is the film thickness. As shown in Eq. (13), ε_f will be highest in the thinnest part of the film. As a result, the surfactant will move in the direction of highest γ and this motion will drag liquid along with it. The latter effect is referred to as the Marangoni effect, which reduces further thinning of the film and hence will reduce coalescence during emulsification.

Another role of the surfactant is to initiate interfacial instability, e.g., by creating turbulence and Rayleigh and Kelvin–Helmholtz instabilities. Turbulence eddies tend to disrupt the interface since they create local pressures. Interfacial instabilities may also occur for cylindrical threads of disperse phase during emulsification. Such cylinders undergo deformation and become unstable under certain conditions. The presence of surfactants will accelerate these instabilities as a result of the interfacial tension gradient.

VI. SURFACTANTS AS DISPERSANTS

Surfactants are used as dispersants for solids in liquid dispersions (suspensions). The latter are prepared by two main procedures, namely, condensation methods (that are based on building up the particles from molecular units) and dispersion methods, whereby larger “lumps” of the insoluble solid are subdivided by mechanical or other methods (referred to as comminution). The condensation methods involve two main processes, nucleation and growth. Nucleation is a spontaneous process of the appearance of a new phase from a metastable (supersaturated) solution of the material in question. The initial stages of nucleation result in the formation of small nuclei where the surface-to-volume ratio is very high and hence the role of specific surface energy is very important. With the progressive increase in the size of nuclei, the ratio becomes lower and eventually larger crystals appear, with a corresponding reduction in the role played by the specific surface energy.

The addition of surfactants, which can either adsorb on the surface of a nucleus or act as a center for inducing nucleation, can be used to control the process of nucleation and the stability of the resulting nuclei. This is due to their effect on the specific surface energy, on the one hand, and their ability to incorporate the material in the micelles, on the other.

Surfactants play a major role in the preparation of suspensions of polymer particles by heterogeneous nucleation. In emulsion polymerization, the monomer is emulsified in a nonsolvent (usually water) using a surfactant, whereas the initiator is dissolved in the continuous phase. The role of surfactants in this process is obvious since nucleation may occur in the swollen surfactant micelle. Indeed, the number of particles formed and their size depend on the nature of surfactant and its concentration (which determines the number of micelles formed).

Dispersion polymerization differs from emulsion polymerization in that the reaction mixture, consisting of monomer, initiator, and solvent (aqueous or nonaqueous), is usually homogeneous. As polymerization proceeds, polymer separates out and the reaction continues in a heterogeneous manner. A polymeric surfactant of the block or graft type (referred to as "protective colloid") is added to stabilize the particles once formed.

The role of surfactants in the preparation of suspensions by dispersion of a powder in a liquid and subsequent wet milling (comminution) can be understood by considering the steps involved in this process. Three steps may be distinguished: wetting of the powder with the liquid, breaking of aggregates, and agglomerates, and comminution. Surfactants play a crucial role in every step. For wetting the powder with the liquid, it is necessary to lower its surface tension and also reduce the solid/liquid interfacial tension by surfactant adsorption. The latter results in reduction of the contact angle of the liquid on the solid substrate.

The work of dispersion, W_d , involved in wetting a unit area of the solid substrate is given by the difference between the interfacial tension of the solid/liquid interface, γ_{SL} , and that of the solid/vapor interface, γ_{SV} ,

$$W_d = \gamma_{SL} - \gamma_{SV}. \quad (14)$$

Using Young's equation,

$$\gamma_{SV} - \gamma_{SL} = \gamma_{LV} \cos \theta, \quad (15)$$

one obtains

$$W_d = -\gamma_{LV} \cos \theta. \quad (16)$$

Thus, the work of dispersion depends on γ_{LV} and θ , both of which are reduced by the addition of surfactant.

Breaking of aggregates (clusters joined at their particle faces) and agglomerates (clusters joined at the corners of the particles) is also aided by the addition of surfactants.

Surfactants also aid the comminution of the particles by bead milling, whereby adsorption of the surfactant at the solid/liquid interface and in "cracks" facilitates their disruption into smaller units.

VII. ROLE OF SURFACTANTS IN STABILIZATION OF EMULSIONS AND SUSPENSIONS

Surfactants are used for stabilization of emulsions and suspensions against flocculation, Ostwald ripening, and coalescence. Flocculation of emulsions and suspensions may occur as a result of van der Waals attraction, unless a repulsive energy is created to prevent the close approach of droplets or particles. The van der Waals attraction G_A between two spherical droplets or particles with radius R and surface-to-surface separation h is given by the Hamaker equation,

$$G_A = -\frac{AR}{12h}, \quad (17)$$

where A is the effective Hamaker constant, which is given by the difference of the sum of all dispersion forces of the particles, A_{11} , and the medium, A_{22} ,

$$A = (A_{11}^{1/2} - A_{22}^{1/2})^2. \quad (18)$$

Equation (17) shows that G_A increases with a decrease in h , and at small distances it can reach very large values (several hundred kT units). To overcome this everlasting attractive force and hence prevent flocculation of the emulsion or suspension, one needs to create a repulsive energy that "shields" the van der Waals energy. Two main types of repulsion may be distinguished. The first is the result of the presence of double layers, as, for example, when using ionic surfactants. The latter become adsorbed on the droplet or particle surface, and this results in the formation of a surface charge (which is characterized by a surface potential ψ_o). This surface charge is neutralized by counterions (which have a sign opposite that of the surface charge) which extend a large distance from the surface (which depends on the electrolyte concentration and valency). Around the particle surface, there will be an unequal distribution of counterions and co-ions (which have the same charge sign as the surface). The surface charge plus the counter- and co-ions form the electrical double layer, which may be characterized by a thickness ($1/\kappa$) that increases with a decrease in electrolyte concentration and valency.

When two droplets or particles approach a distance h that is smaller than twice the double-layer thickness, repulsion occurs due to double-layer overlap (the double layers on the two particles cannot develop completely).

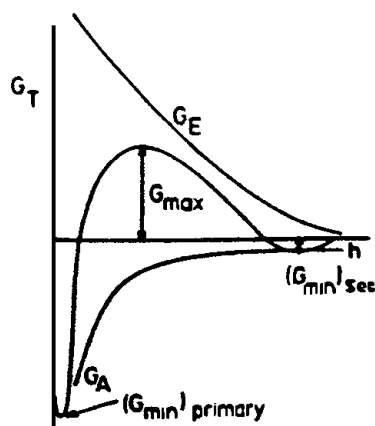


FIGURE 9 Form of the interaction energy–distance curve according to the DLVO theory.

The electrostatic energy of repulsion is given by the expression

$$G_E = 2\pi R \varepsilon_r \varepsilon_0 \Psi_0^2 \ln[1 + \exp(-\kappa h)], \quad (19)$$

where ε_r is the relative permittivity and ε_0 is the permittivity of free space.

It is clear from Eq. (19) that G_E increases with an increase in Ψ_0 (or zeta potential) and a decrease in κ (i.e., a decrease in electrolyte concentration and valency). Combination of G_E and G_A forms the basis of the stability of lyophobic colloids proposed by Deryaguin and Landau and Verwey and Overbeek, referred to as the DLVO theory. The energy–distance curve based on the DLVO theory is represented schematically in Fig. 9. It shows two minima, at long and short distances, $(G_{\min})_{\text{sec}}$ and $(G_{\min})_{\text{primary}}$ respectively, and an energy maximum $G_{\max}$ at intermediate distances. If $G_{\max}$ is high (>25 kT) the energy barrier prevents close approach of the droplets or particles, and hence irreversible flocculation into the primary minimum is prevented. This high-energy barrier is maintained at low electrolyte concentrations ($<10^{-3}$ mol dm $^{-3}$) and high surface (or zeta) potentials.

The second repulsive energy (referred to as steric repulsion) is produced by the presence of adsorbed surfactant layers of nonionic surfactants, such as alcohol ethoxylates or A–B, A–B–A block, or BA $_n$ graft copolymers, where B is the “anchor” chain and A is the stabilizing chain [mostly based on polyethylene oxide (PEO) for aqueous systems]. When two droplets or particles with adsorbed PEO chains of thickness δ approach a separation distance h such that $h < 2\delta$, repulsion occurs as a result of two main effects. The first arises as a result of the unfavorable mixing of the PEO chains, when these are in good solvent conditions. This is referred to as G_{mix} and is given by the following expression:

$$\frac{G_{\text{mix}}}{KT} = \frac{4\pi\phi_2^2}{3V_1} \left(\frac{1}{2} - \chi \right) \left(\delta - \frac{h}{2} \right)^2 \left(3R + 2\delta + \frac{h}{2} \right), \quad (20)$$

where ϕ_2 is the volume fraction of the chains in the adsorbed layer, V_1 is the molar volume of the solvent, and χ is the chain–solvent (Flory–Huggins) interaction parameter. It is clear from Eq. (20) that when $\chi < 0.5$ (i.e., the chains are in good solvent conditions), G_{mix} is positive and the interaction is repulsive. In contrast, when $\chi > 0.5$ (i.e., the chains are in poor solvent conditions), G_{mix} is negative and the interaction is attractive. The condition $\chi = 0.5$, referred to as the θ -point, represents the onset of flocculation.

The second contribution to the steric interaction arises from the loss of configurational entropy of the chains on significant overlap. This effect is referred to as entropic, volume restriction, or elastic interaction, G_{el} . The latter increases very sharply with a decrease in h when the latter is less than δ . A schematic representation of the variation of G_{mix} , G_{el} , G_A , and G_T ($=G_{\text{mix}} + G_{\text{el}} + G_A$) is given in Fig. 10. The total energy–distance curve shows only one minimum, at $h \sim 2\delta$, the depth of which depends on δ , R , and A . At a given R and A , $G_{\min}$ decreases with an increase in δ . With small particles and thick adsorbed layers ($\delta > 5$ nm), $G_{\min}$ becomes very small ($<kT$) and the dispersion approaches thermodynamic stability. This shows the importance of steric stabilization in controlling the flocculation of emulsions and suspensions.

Surfactants are also used for reduction of Ostwald ripening. The latter arises from the difference in solubility between small and large particles. The smaller particles will have a higher solubility compared with larger ones. This is the result of the higher radius of curvature of smaller particles (note that the solubility of any droplet or particle S is inversely proportional to its radius R ; $S = 2\gamma/R$). With time, smaller droplets or particles dissolve and their molecules diffuse and become deposited on larger droplets

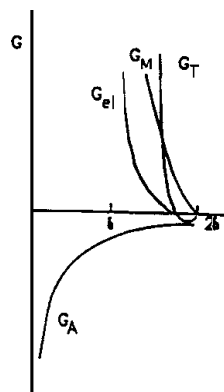


FIGURE 10 Variation of G_{mix} , G_{el} , G_A , and G_T with h for a sterically stabilized dispersion.

or particles. Surfactants reduce Ostwald ripening by two main mechanisms. First, by reduction of interfacial tension, the rate of Ostwald ripening is reduced. Second, as a result of interfacial tension gradients, the Gibbs elasticity causes a significant reduction of Ostwald ripening. The most effective surfactants are those with strong adsorption at the interface.

Surfactants also reduce the coalescence of emulsion droplets. The latter process occurs as a result of thinning and disruption of the liquid film between the droplets on their close approach. The latter causes surface fluctuations, which may increase in amplitude and the film may collapse at the thinnest part. This process is prevented by the presence of surfactants at the O/W interface, which reduce the fluctuations as a result of the Gibbs elasticity and/or interfacial viscosity. In addition, the strong repulsion between the surfactant layers (which could be electrostatic and/or steric) prevents close approach of the droplets, and this reduces any film fluctuations. In addition, surfactants may form multilayers at the O/W interface (lamellar liquid crystalline structures), and this prevents coalescence of the droplets.

VIII. ROLE OF SURFACTANTS IN SOLUBILIZATION AND MICROEMULSIONS

A. Solubilization

Solubilization is the formation of a thermodynamically stable, isotropic solution of a substance (the solubilize), normally insoluble or slightly soluble in water, by the addition of a surfactant (the solubilizer). The micelles of the surfactant cause solubilization of the substrate, producing an isotropic solution of the chemical. The solubilize can be incorporated in the surfactant micelle in different ways, depending on the nature of the substrate and the surfactant micelles. For hydrophobic substrates, the molecules become incorporated in the hydrocarbon core of the micelle. With more polar substrates, the molecules may become incorporated in the hydrophilic PEO chains of the micelle or they may be simply adsorbed at the micelle surface.

Solubilization is applied in many industrial processes for the administration of insoluble chemicals, e.g., in dyeing, in drug administration, and in agrochemical applications. The process of solubilization is also important in detergency, whereby fats and oils are removed by incorporation into the hydrocarbon core of the micelles.

B. Microemulsions

Microemulsions are isotropic systems consisting of oil, water, and surfactant(s) which are stable in the thermo-

dynamic sense, consisting of nearly isodisperse, small droplets (usually in the range of 5–50 nm) in another liquid. The small size of the droplets results in the transparent or translucent appearance of microemulsion.

Several theories have been proposed to account for the thermodynamic stability of microemulsions. The most recent theories showed that the driving force for microemulsion formation is the ultralow interfacial tension (in the region of 10^{-4} – 10^{-2} mN m⁻¹). This means that the energy required for formation of the interface (the large number of small droplets) $\Delta A\gamma$ is compensated by the entropy of dispersion $-T\Delta S$, which means that the free energy of formation of microemulsions ΔG is zero or negative.

The ultralow interfacial tension can be produced by using a combination of two surfactants, one predominantly water soluble (such as sodium dodecyl sulfate) and the other predominantly oil soluble (such as a medium-chain alcohol, e.g., pentanol or hexanol). In some cases, one surfactant may be sufficient to produce the microemulsion, e.g., Aerosol OT (dioctyl sulfosuccinate), which can produce a W/O microemulsions. Nonionic surfactants, such as alcohol ethoxylates, can also produce O/W microemulsions, within a narrow temperature range. As the temperature of the system increases, the interfacial tension decreases, reaching a very low value near the phase inversion temperature. At such temperatures, an O/W microemulsion may be produced.

Microemulsions have attracted considerable attention for application in industry. In the early days of their discovery, microemulsions were used in the leather industry, cutting oils, dry cleaning, flavorings, agrochemicals, and pharmaceuticals. However, the main potential application of microemulsions will be in tertiary oil recovery and as reaction media for enzymes and production of nanoparticles (e.g., for application in the electronic industry). Another application of microemulsions is in the field of solar energy production, e.g., production of hydrogen by decomposition of water using UV light.

IX. SURFACTANTS IN FOAMS

The role of surfactants in stabilization/destabilization of foam (air/liquid dispersions) is similar to that for emulsions. This is due to the fact that foam stability/instability is determined by the surface forces operative in liquid films between air bubbles. In many industrial applications, it is essential to stabilize foams against collapse, e.g., with many food products, foam in beer, fire-fighting foam, and polyurethane foams that are used for furniture and insulation. In other applications, it is essential to have an effective way of breaking the foam, e.g., in distillation

columns, crude oils, and effluent streams. In the case where foam stability is desirable, it is essential to choose surfactants that enhance the Gibbs–Marangoni effect and produce a viscoelastic film that provides a mechanical barrier preventing foam collapse. This explains the application of protein film for fire-fighting foams. A particularly important process in foam formation and stabilization is Ostwald ripening, which results from gas diffusion from smaller air bubbles to larger ones. This is the result of the higher Laplace pressure of smaller air bubbles. As mentioned in the section on stabilization of emulsions, this process is opposed by the reduction of the interfacial tension and creation of an interfacial tension gradient (Gibbs elasticity).

In the case where foam instability is desirable, it is essential to choose surfactants that weaken the Gibbs–Marangoni effect. A more surface-active material such as a poly(alkyl) siloxane is added to destabilize the foam. The siloxane surfactant adsorbs preferentially at the air/liquid interface, thus displacing the original surfactant that stabilizes the foam. In many cases, the siloxane surfactant is produced as an emulsion which also contains hydrophobic silica particles. This combination produces a synergetic effect for foam breaking.

X. SURFACTANTS IN WETTING PHENOMENA

Wetting is important in many industrial systems, e.g., mineral flotation, detergency, crop protection, dispersion of powders in liquids, and coatings. When a drop of a liquid is placed on a solid surface, the liquid either spreads, forming a thin uniform film (complete wetting with zero contact angle θ), or remains as a discrete droplet with a measurable contact angle (partial wetting). The value of the contact angle is used as a measure of wetting: when $\theta = 0^\circ$, complete wetting occurs; when $\theta = 180^\circ$, the surface is described as nonwetable. When $\theta < 90^\circ$, the surface is described as being partially wetted, whereas when $\theta > 90^\circ$ the surface is described as being poorly wetted by the liquid. Thus, to enhance the wetting of an aqueous solution on a hydrophobic substrate, one adds a surfactant, which lowers the surface tension of water and adsorbs on the hydrophobic substrate in a specific manner, i.e., with the hydrophobic alkyl chain being attached to the substrate, leaving the polar head group in the aqueous medium. In contrast, to reduce the wetting of an aqueous solution on a hydrophilic surface (e.g., in waterproofing), one adds a surfactant with the opposite orientation, i.e., the polar head group being attached to the surface, leaving the hydrophobic alkyl chain pointing to the aqueous medium. An example of the latter process is

waterproofing of fabrics, whereby a cationic surfactant is sometimes used. The positive head group of the surfactant is attached to the negative charges on the fabric, leaving the hydrophobic alkyl chains pointing to the solution. The same process applies for fabric softeners, which usually consist of dialkyl quaternary ammonium surfactants.

XI. APPLICATION OF SURFACTANTS IN COSMETICS AND PERSONAL CARE PRODUCTS

Cosmetic and personal care products are designed to deliver a functional benefit and to enhance the psychological well-being of consumers by increasing their aesthetic appeal. Many cosmetic and personal care formulations are designed to clean hair, skin, etc., and impart a pleasant odor, make the skin feel smooth, provide moisturizing agents, provide protection against sunburn, etc. Most cosmetic and personal care products consist of complex systems of emulsions, creams, lotions, suspoemulsions (mixtures of emulsions and suspensions), multiple emulsions, etc. All these complex systems consist of several components of oil, water, surfactants, coloring agents, fragrances, preservatives, vitamins, etc. The role of surfactants in these complex formulations is crucial in designing the system, in achieving long-term physical stability and the required “skin-feel” on application. Conventional surfactants of the anionic, cationic, amphoteric, and nonionic types are used in cosmetics and personal care applications. These surfactants may not cause any adverse toxic effects. Besides the synthetic surfactants used in the preparation of systems such as emulsions, creams, lotions, and suspensions, several other naturally occurring materials have been introduced and there is a trend in recent years to use such natural products in the belief that they are safer for application. Several synthetic surfactants that are applied in cosmetics and personal care products may be listed, such as carboxylates, ether sulfates, sulfates, sulfonates, quaternary amines, betaines, and sarcosinates. The ethoxylated surfactants are probably the most widely used surfactants in cosmetics. Being uncharged, these molecules have a low skin sensitization potential. This is due to their low binding to proteins. Unfortunately, these nonionic surfactants are not the most friendly materials to produce (the ethoxylation process is rather dangerous), and one has to ensure a very low level of free ethylene oxide, which may form dioxane (that is carcinogenic) on storage. Another problem with ethoxylated surfactants is their degradation by oxidation or photooxidation processes. These problems are reduced by using sucrose esters obtained by esterification of the sugar hydroxyl group with fatty acids such as lauric and stearic acid. In this case, the

problem of contamination is reduced and the surfactants are still mild to the skin since they do not interact with proteins.

Another class of surfactants that are used in cosmetics and personal care products is the phosphoric acid esters. These molecules are similar to the phospholipids that are the building blocks of the stratum corneum (the top layer of the skin, which is the main barrier for water loss). Glycerine esters, in particular, triglycerides, are also frequently used. Macromolecular surfactants of the A–B–A block type [where A is PEO and B is polypropylene oxide (PPO)] are also frequently used in cosmetics. Another important naturally occurring class of polymeric surfactants is the proteins, which can be used effectively as emulsifiers.

In recent years, there has been a great trend toward using volatile silicone oils in many cosmetic formulations. Due to their low surface energy, silicone oils help spread the various active ingredients over the surface of the skin, hair, etc. While many silicone oils can be emulsified using conventional hydrocarbon surfactants, several silicone-type surfactants have been introduced for their effective emulsification and long-term stability. These silicone surfactants consist of a methyl siloxane backbone with pendent groups of PEO and PPO. These polymeric surfactants act as steric stabilizers.

XII. APPLICATION OF SURFACTANTS IN PHARMACEUTICALS

Surfactants play an important role in pharmaceutical formulations. A large number of drugs are surface active, e.g., chlorpromazine, diphenyl methane derivatives, and tricyclic antidepressants. The solution properties of these surface-active drugs play an important role in their biological efficacy. Surface-active drugs tend to bind hydrophobically to proteins and other biological macromolecules. They tend to associate with other amphipathic molecules such as other drugs, bile salts, and receptors. Many surface-active drugs produce intralysosomal accumulation of phospholipids which are observable as multilamellar objects within the cell. The interaction between surfactant drug molecules and phospholipid renders the phospholipid resistant to degradation by lysosomal enzymes, resulting in their accumulation in the cell.

Many local anesthetics have significant surface activity and it is tempting to correlate such surface activity with their action. Other important factors such as partitioning of the drug into the nerve membrane may also play an important role. Accumulation of drug molecules in certain sites may allow them to reach concentrations whereby micelles are produced. Such aggregate units may cause significant biological effects.

Several naturally occurring amphipathic molecules (in the body) exist, such as bile salts, phospholipids, and cholesterol, which play an important role in various biological processes. Their interactions with other solutes, such as drug molecules, and with membranes are also very important. The most important surface-active species in the body are the phospholipids, e.g., phosphatidylcholine (lecithin). These lipids (which may be produced from egg yolk) are used as emulsifiers for many intravenous formulations, such as fat emulsions and anesthetics. Lipids can also be used to produce liposomes and vesicles which can be applied for drug delivery. When dispersed into water, they produce lamellar structures, which then produce multilamellar spherical units (liposomes). On sonication of these multilamellar structures, single spherical bilayers or vesicles (10–40 nm) are produced. Both lipid-soluble and water-soluble drugs can be entrapped in the liposomes. Liposoluble drugs are solubilized in the hydrocarbon interiors of the lipid bilayers, whereas water-soluble drugs are intercalated in the aqueous layers.

One of the most important application of surfactants in pharmacy is to solubilize insoluble drugs. Several factors may be listed that influence solubilization such as the surfactant and solubilizate structure, temperature, and added electrolyte or nonelectrolyte. Solubilization in surfactant solutions above the cmc offers an approach to the formulation of poorly insoluble drugs. Unfortunately, this approach has some limitations, namely, the finite capacity of the micelles for the drug, the possible short- or long-term adverse effects of the surfactant on the body, and the concomitant solubilization of other ingredients such as preservatives and flavoring and coloring agents in the formulation. Nevertheless, there is certainly a need for solubilizing agents for increasing the bioavailability of poorly soluble drugs. The use of cosolvents and surfactants to solve the problem of poor solubility has the advantage that the drug entity can be used without chemical modification and toxicological data on the drug may not be repeated.

Surfactants are also used for general formulation of drugs, e.g., as emulsifying agents, dispersants for suspensions, and wetting agents for tablets. Surfactant molecules incorporated in the formulation can affect drug availability in several ways. The surfactant may influence the disintegration and dissolution of solid dosage forms or control the rate of precipitation of drugs administered in solution form, by increasing the membrane permeability and affecting membrane integrity. Release of poorly soluble drugs from tablets and capsules for oral use may be increased by the presence of surfactants, which may decrease the aggregation of drug particles and, therefore, increase the area of the particles available for dissolution. The lowering of surface tension may also be a factor in aiding the penetration of water into the drug mass. Above

the cmc, the increase in solubilization can result in more rapid rates of drug dissolution.

XIII. APPLICATION OF SURFACTANTS IN AGROCHEMICALS

Besides the use of surfactants for formulation of all agrochemical formulations (suspensions, emulsions, microemulsion, microcapsules, water-dispersible grains, granules, etc.), these molecules play a major role in optimization of biological efficacy. This can be understood if one considers the steps during application of the crop spray, which involve a number of interfaces. The first interface during application is that between the spray solution and the atmosphere (air), which governs the droplet spectrum, rate of evaporation, drift, etc. In this respect, the rate of adsorption of the surfactant molecules at the air/liquid interface is of vital importance. In a spraying process a fresh liquid surface is continuously being formed. The surface tension of this liquid (referred to as the dynamic surface tension) depends on the relative ratio between the time taken to form an interface and the rate of adsorption of the surfactant from the bulk solution to the air/liquid interface, which depends on the rate of diffusion of the surfactant molecule. The rate of diffusion is directly proportional to the diffusion coefficient, D , of the molecule (which is inversely proportional to its radius) and the surfactant concentration. Thus, for effective lowering of the dynamic surface tension during a spraying process, one needs surfactants with a high D and sufficiently high concentrations. However, the actual situation is not simple since one has an equilibrium between surfactant micelles and monomers. The latter diffuse to the interface and become adsorbed, and hence the equilibrium between micelles and monomers is disturbed. Surfactant micelles then break to supply monomers in the bulk. Thus, the dynamic surface tension also depends on the lifetime of the micelle.

Surfactants also have a large influence on spray impaction and adhesion, which is very important for maximizing capture of the drops by the target. For adhesion to take place, the difference in surface energy of the droplet in flight, $E_o (=4\pi R^2\gamma)$, and that at the surface, E_s (which depends on the contact angle, θ , of the drop on the substrate), should balance the kinetic energy of the drop ($\frac{1}{2}mv^2$, where m is the mass of the drop and v its velocity). For adhesion to occur, $E_o - E_s > \frac{1}{2}mv^2$. Surfactants clearly enhance adhesion by lowering γ and θ .

Surfactants also play a major role in reducing droplet sliding and increasing spray retention. When a drop impinges on an inclined surface (such as a leaf surface), it starts to slide as a result of gravity. During this process,

the droplet produces an advancing contact angle θ_A and a receding contact angle θ_R . The latter is lower than the former, and the difference between the two angles ($\theta_A - \theta_R$) is referred to as contact angle hysteresis. As a result of this sliding process, an area of the surface becomes dewetted (at the back) and an equal area becomes wetted at the front. When the difference between the work of dewetting and that of wetting (which is determined by the contact angle hysteresis) balances the gravity force, sliding stops and the droplet stays retained on the surface. Thus, surfactants which affect the surface tension of the liquid and give this contact angle hysteresis reduce drop sliding and enhance spray retention.

Another role of surfactants in crop sprays is to enhance the wetting and spreading of the droplets on the target surface. This process governs the final distribution of the agrochemical over the area to be protected. The optimum degree of coverage in any spray application depends on the mode of action of the agrochemical and the nature of the pest to be controlled. On evaporation of the drops, deposits are produced whose nature depends on the nature of the surfactant and interaction with the agrochemical molecules or particles. These deposits may contain liquid crystalline phases when the surfactant concentration reaches high values. In many cases, long-lasting deposits are required to ensure supply of the agrochemical, e.g., with systemic fungicides. These deposits may enhance the tenacity of the agrochemical on the leaf surface and hence they enhance rain-fastness.

Finally, surfactants may have a direct effect on the biological efficacy by enhancing the penetration of agrochemical molecules through various barriers, such as plant cuticle and various other membranes. This enhanced penetration may be caused by solubilization of the active ingredient by the surfactant micelles. The latter may enhance flux of the chemical through the plant by increasing the concentration gradient at the interface.

XIV. APPLICATION OF SURFACTANTS IN THE FOOD INDUSTRY

The use of surfactants in the food industry has been known for centuries. Naturally occurring surfactants such as lecithin from egg yolk or soybean and various proteins from milk are used for the preparation of many food products, such as mayonnaise, salad creams, dressing, and desserts. Polar lipids such as monoglycerides have been introduced as emulsifiers for food products. More recently, synthetic surfactants such as sorbitan esters (Spans) and their ethoxylates (Tweens), sucrose esters, have been used in food emulsions. It should be mentioned that the structures of many food emulsions is complex, and in

many cases several phases may exist. Such structures may exist under nonequilibrium conditions and the state of the system may depend to a large extent on the process used for preparing the system, its prehistory, and the conditions to which it is subjected.

Food grade surfactants are, in general, not soluble in water, but they can form association structures in aqueous medium that are liquid crystalline in nature. These liquid crystalline structures are produced by heating the solid emulsifier (which is dispersed in water) to a temperature above its Krafft temperature. On cooling such a system, a "gel" phase is produced which becomes incorporated with the emulsion droplets. These gel phases produce the right consistency for many food emulsions.

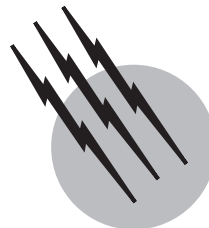
Proteins, which are also surface active, can be used to prepare food emulsions. The protein molecules adsorb at the O/W interface and they may remain in their native state (forming a "rigid" layer of unfolded molecules) or undergo unfolding, forming loops, tails, and trains. These protein molecules stabilize the emulsion droplets, either by a steric stabilization mechanism or by producing a mechanical barrier at the O/W interface.

SEE ALSO THE FOLLOWING ARTICLES

CHEMICAL THERMODYNAMICS • MESOPOROUS MATERIALS, SYNTHESIS AND PROPERTIES • MICELLES • SILICONE (SILOXANE) SURFACTANTS

BIBLIOGRAPHY

- Tadros, Th. F. (1984). "Surfactants," Academic Press, London.
- McCutchen (published annually). "Detergents and Emulsifiers," Allied, NJ.
- van Os, N. M., Haak, J. R., and Rupert, L. A. (1993). "Physico-Chemical Properties of Selected Anionic, Cationic and Nonionic Surfactants," Elsevier, Amsterdam.
- Porter, M. R. (1991). "Handbook of Surfactants," Chapman and Hall, London.
- Tadros, Th. F. (1999). In "Principles of Polymer Science and Technology in Cosmetics and Personal Care" (E. D. Goddard and J. V. Gruber, eds.), Chap. 3, Marcel Dekker, New York.
- Griffin, W. C. (1954). *J. Cosmet. Chem.* **5**, 249.
- Lindman, B. (1984). In "Surfactants" (Th. F. Tadros, ed.), Academic Press, London.
- Mukerjee, P., and Mysels, K. J. (1971). "Critical Micelle Concentrations of Aqueous Surfactant Systems," National Bureau of Standards, Washington, DC.
- Hartley, G. S. (1936). "Aqueous Solutions of Paraffin Chain Salts" (Hermann and Cie, Paris).
- Debye, P., and Anaker, E. W. (1951). *J. Phys. Colloid Chem.* **55**, 644.
- McBain, J. W. (1950). "Colloid Science," Heath, Boston.
- Clunies, J. S., Goodman, J. F., and Symons, P. C. (1969). *Trans Faraday Soc.* **65**, 287.
- Rosevaar, F. B. (1968). *J. Soc. Cosmet. Chem.* **19**, 581.
- Anaïsson, E. A. G., and Wall, S. N. (1974). *J. Phys. Chem.* **78**, 1024; (1975) **79**, 857.
- Tanford, C. (1980). "The Hydrophobic Effect," 2nd ed., Wiley, New York.
- Gibbs, J. W. (1928). "Collected Works," Vol. 1, Longman, New York.
- Hough, D. B., and Randall, H. M. (1983). In "Adsorption from Solution at the Solid/Liquid Interface" (G. D. Parfitt and C. H. Rochester, eds.), p. 247, Academic Press, London.
- Clunie, J. S., and Ingram, B. T. (1983). In "Adsorption from Solution at the Solid/Liquid Interface," (G. D. Parfitt and C. H. Rochester, eds.), p. 105, Academic Press, London.
- Walstra, P. (1980). In "Encyclopedia of Emulsion Technology" (P. Becher, ed.), Chap. 2, Marcel Dekker, New York.
- Davies, J. T. (1972). "Turbulence Phenomenon," Chaps. 8–10. Academic Press, New York.
- Chandrasekhar, S. (1961). "Hydrodynamics and Hydrodynamic Instability," Chaps. 10–12, Cleeverdon, Oxford.
- Gibbs, J. W. (1906). "Scientific Papers," Vol. 1, Longman Green, London.
- Volmer, M. (1939). "Kinetic der Phase Bildung," Steinkopf, Dresden.
- Blakely, D. (1975). "Emulsion Polymerization," Applied Science, London.
- Barrett, K. E. J. (1975). "Dispersion Polymerization in Organic Media," John Wiley and Sons, London.
- Hamaker, H. C. (1937). *Physica (Utrecht)* **4**, 1058.
- Deryaguin, B. V., and Landau, L. (1939). *Acta Phy. Chem. USSR* **10**, 33.
- Verwey, E. J., and Overbeek, J. Th. G. (1948). "Theory of Stability of Lyophobic Colloids," Elsevier, Amsterdam.
- Napper, D. H. (1983). "Polymeric Stabilization of Colloidal Dispersions," Academic Press, London.
- Danielsson, I., and Lindman, B. (1981). *Colloids Surf.* **3**, 391.
- Overbeek, J. Th. G. (1978). *Faraday Disc. Chem. Soc.* **65**, 7.
- Overbeek, J. Th. G., de Bruyn, P. L., and Verhoecks, F. (1984). In "Surfactants" (Th. F. Tadros, ed.), p. 111, Academic Press, London.
- Breuer, M. M. (1985). In "Encyclopedia of Emulsion Technology" (P. Becher, ed.), Vol. 2, Chap. 7, Marcel Dekker, New York.
- Attwood, D., and Florence, A. T. (1983). "Surfactant Systems, Their Chemistry, Pharmacy and Biology," Chapman and Hall, New York.
- Tadros, Th. F. (1987). *Aspects Appl. Biol.* **14**, 1.
- Tadros, Th. F. (1994). "Surfactants in Agrochemicals," Marcel Dekker, New York.
- Krog, N. J., and Riisom, T. H. (1985). In "Encyclopedia of Emulsion Technology" (P. Becher, ed.), Vol. 2, p. 321, Marcel Dekker, New York.



Synthetic Fuels

Ronald F. Probst

R. Edwin Hicks

Massachusetts Institute of Technology

- I. Coal, Oil Shale, and Tar Sand Conversion
- II. Thermal Conversion Processes
- III. Technologies
- IV. Biomass Conversion
- V. Outlook

GLOSSARY

Biomass Any material directly or indirectly derived from plant life that is renewable in time periods of less than about 100 years.

Coal Solid fossil hydrocarbon typically composed of from 65 to 75 mass% carbon and about 5 mass% hydrogen, with the remainder oxygen, ash, and smaller quantities of sulfur and nitrogen.

Coproducting Processing of coal and oil simultaneously with the objective of liquefying the coal and upgrading the oil.

Direct hydrogenation Exposure of a carbonaceous raw material to hydrogen at a high pressure.

Direct liquefaction Hydrogenation of a carbonaceous material, usually coal, to form a liquid fuel by direct hydrogen addition in the presence of a catalyst or by transfer of hydrogen from a solvent.

Gasification Conversion of a carbonaceous material into a gas, with the principal method to react steam with coal in the presence of air or oxygen in a vessel called a gasifier.

Hydrotreating Catalytic addition of hydrogen to liquid

fuels to remove oxygen, nitrogen, and sulfur and to make lighter fuels by increasing the hydrogen-to-carbon ratio.

Indirect hydrogenation Reaction of a carbonaceous raw material with steam, with the hydrogen generated within the system.

Indirect liquefaction Combination of a synthesis gas composed of carbon monoxide and hydrogen over a suitable catalyst to form liquid products such as gasoline and methanol.

Oil shale Sedimentary rock containing kerogen, a high molecular mass hydrocarbon, that is insoluble in common solvents and is not a member of the petroleum family.

Pyrolysis Reduction of the carbon content in a raw hydrocarbon by distilling volatile components to yield solid carbon, as well as gases and liquids with a higher hydrogen fraction than the original material.

Reactor Vessel used for gasification, liquefaction, and pyrolysis, with the three main types the moving packed bed, the entrained flow, and the fluidized bed reactor.

Retorting Pyrolysis of oil shale to produce oil in a vessel called a retort.

SNG Substitute natural gas that consists primarily of methane manufactured mainly by the catalytic synthesis of carbon monoxide and hydrogen.

Synthesis Combination of a gas whose major active components are carbon monoxide and hydrogen over a suitable catalyst to form a large number of products including methane, methanol, gasoline, and alcohols.

Synthetic fuels Gaseous or liquid fuels manufactured by hydrogenating a naturally occurring carbonaceous raw material or by removing carbon from the material.

Tar sands Mixture of sand grains, water, and a high-viscosity hydrocarbon called bitumen, which is a member of the petroleum family.

SYNTHETIC FUELS may be gaseous, liquid, or solid and are obtained by converting a carbonaceous material to another form. The most abundant naturally occurring materials for producing synthetic fuels are coal, oil shale, tar sands, and biomass. The conversion of these materials is undertaken to provide synthetic gas or oil to replace depleted or unavailable natural resources and also to remove sulfur or nitrogen, which, when burned, gives rise to undesirable air pollutants. The manufacture of synthetic fuels can be regarded as a process of hydrogenation since common fuels have a higher hydrogen content than the raw materials. All synthetic fuel processes require an energy input to accomplish the conversion. Most of the thermal conversion processes are applicable to all carbonaceous materials. The biochemical processes of fermentation and biological decomposition are specific to biomass.

I. COAL, OIL SHALE, AND TAR SAND CONVERSION

A. Synthetic Fuel Manufacture and Properties

To manufacture synthetic fuels, hydrogenation of the naturally occurring raw materials, or carbon removal, is usually required since common fuels such as gasoline and natural gas have a higher hydrogen content than the raw materials. The source of the hydrogen that is added is water. A typical bituminous coal has a carbon-to-hydrogen mass ratio of about 15, while methane, which is the principal constituent of natural gas, has a carbon-to-hydrogen mass ratio of 3. In between, the corresponding ratio for crude oil is about 9, and that for gasoline 6.

The organic material in both tar sands and high-grade oil shale has a carbon-to-hydrogen mass ratio of about 8, which is close to that of crude oil. However, the mineral content of rich tar sands in the form of sand or sandstone is about 85 mass%, and that of high-grade oil shale, in the form of sedimentary rock, is about the same. Therefore, very large volumes of solids must be handled to recover

relatively small quantities of organic matter from oil shale and tar sands. On the other hand, the mineral content of coal in the United States averages about 10% by mass.

In any conversion to produce a fuel of a lower carbon-to-hydrogen ratio, the hydrogenation of the raw fossil fuel may be direct, indirect, or by pyrolysis, either alone or in combination. Direct hydrogenation involves exposing the raw material to hydrogen at a high pressure. Indirect hydrogenation involves reacting the raw material with steam, with the hydrogen generated within the system. In pyrolysis the carbon content is reduced by heating the raw hydrocarbon until it thermally decomposes, distilling off the volatile components to yield solid carbon, together with gases and liquids having higher fractions of hydrogen than the original material.

Fuels that will burn cleanly require that sulfur and nitrogen compounds be removed from the gaseous, liquid, and solid products. As a result of the hydrogenation process, sulfur and nitrogen, which are always present to some degree in the original raw fossil fuel, are reduced to hydrogen sulfide and ammonia, respectively. Hydrogen sulfide and ammonia are present in the gas made from coal or released during the pyrolysis of oil shale and tar sands and, also, are present in the gas generated in the hydrotreatment of pyrolysis oils and synthetic crude oils.

Synthetic fuels include liquid fuels such as fuel oil, diesel oil, gasoline, and methanol, clean solid fuels, and low-calorific value, medium-calorific value, and high-calorific value gas. The gas is referred to here as low-CV, medium-CV, and high-CV gas, respectively. In British units, which are still used interchangeably, the corresponding reference is to low-Btu, medium-Btu, and high-Btu gas. Low-CV gas, often called producer or power gas, has a calorific value of about 3.5 to 10 million joules per cubic meter (MJ/m^3) or, in British units, 90 to 270 British thermal units per standard cubic foot (Btu/scf). This gas is an ideal turbine fuel. Medium-CV gas is loosely defined as having a calorific value of about 10 to 20 MJ/m^3 (270–540 Btu/scf). This gas is also termed power gas and, sometimes, industrial gas, as well as synthesis gas. It may be used as a fuel gas, as a source of hydrogen for direct liquefaction, or for the synthesis of methanol and other liquid fuels. Medium-CV gas may also be used for the production of high-CV gas, which has a calorific value in the range of about 35–38 MJ/m^3 (940–1020 Btu/scf) and is normally composed of more than 90% methane. This gas is a substitute for natural gas and suitable for economic pipeline transport. For these reasons it is referred to as substitute natural gas (SNG) or pipeline gas.

B. History

Synthetic fuel manufacture, although often thought of as a modern technology, is not new, nor has it been limited

in the past to small-scale development. What is different today is the increased fundamental chemical and physical understanding of the complex conversion processes that is built into the technologies, the application of modern engineering, and systems designed to ensure environmentally sound operation. What is not different is the history of synthetic fuel manufacture, whose on-again, off-again commercialization since the beginning of the nineteenth century has been buffeted by the real or perceived supply of natural resources of oil and gas. In the late 1970s, following the Arab oil embargo of 1973, worldwide commercial synthetic fuel manufacture appeared to be on the verge of reality. By the 1990s, however, there seemed scant likelihood for this to take place in the twentieth century, with most, though not all, work in the field reduced to a small research and development level. Historical evidence, however, indicates that any prediction of full-scale commercialization is at best risky and more likely unreliable.

As early as 1792, Murdoch, a Scottish engineer, distilled coal in an iron retort and lit his home with the coal gas produced. By the early part of the nineteenth century, gas manufactured by the distillation of coal was introduced for street lighting, first in London in 1812, following which its use for this purpose spread rapidly throughout the major cities of the world. This coal gas contained about 50% hydrogen and from 20 to 30% methane, with the remainder principally carbon monoxide. Its calorific value was about 19 MJ/m³ (500 Btu/scf), and this value served as the benchmark for the “town gas” industry. In the latter part of the nineteenth century, gasification technologies, employing the reaction of air and steam with coal, were developed and the use of “synthetic” gas for domestic and industrial application became widespread. Commercial “gas producers” yielded a gas with a low calorific value of about 5–6.5 MJ/m³ (130–160 Btu/scf) and were used on-site to produce gas for industrial heating. In the early part of the twentieth century the availability of natural gas with a calorific value of 37 MJ/m³ (1000 Btu/scf) began to displace the manufactured gas industry, which, subsequent to the end of World War II, virtually disappeared worldwide.

Following the Arab oil embargo of 1973, construction of a number of commercial-scale coal conversion plants was undertaken in the United States to produce SNG on scales up to 7 million m³/day (250 million scf/day). The largest project to be completed was the Great Plains coal gasification plant in North Dakota, which has a design capacity of 3.9 million m³/day (138 million scf/day) of SNG. The plant started up in 1984 and was operated in turn by the U.S. Department of Energy and the Dakota Gasification Company. Production rates have increased beyond the design rate and, by 1991, had reached 4.5 million m³/day (160 million scf/day) of SNG. Several smaller coal gasi-

fication processes remained in operation. The Cool Water plant in California, which shut down at the beginning of 1989, manufactured 2 million m³/day (72 million scf/day) of 9 MJ/m³ (250 Btu/scf) gas. The gas was used to produce about 100 MW (net) of electric power in combustion and steam turbine generators. The plant was reopened in 1993 using a mixed feed of coal and sewage sludge.

The history of coal liquefaction is considerably more recent than that of coal gasification. Direct liquefaction, in which the coal is exposed to hydrogen at a high pressure, can be traced to the work of Bergius in Germany from 1912 to 1926. Commercial-size hydrogenation units for the production of motor fuels began in Germany in 1926, and by 1939 the output was estimated to be 4 million liters of gasoline per day. Liquid fuel production and, in particular, oil production are most frequently quoted in barrels per day, where 1 barrel (bbl) equals 42 U.S. gal or about 160 liters. The German production of gasoline was therefore about 250,000 bbl/day. During World War II this direct liquefaction production from some 12 plants expanded to about 1 million bbl/day of gasoline. Activity in direct coal liquefaction paralleled that in gasification. Most of the work in the 1970s and early 1980s centered about the development of second-generation plants to run at lower pressures of from 10 to 20 million pascals (MPa). The original German Bergius-type units were run at from 25 to 70 MPa. It may be noted that 1 million Pa is about 10 atm or 147 lb/in.². The larger of the pilot plants in the United States and Germany operated at about 200 to 250 ton-per-day coal feeds, equal to about 550 to 700 bbl/day of synthetic crude oil output. Throughout this chapter, ton (t) refers to the unit in use with SI units.

The principal method of indirect liquefaction is to react carbon monoxide and hydrogen produced by coal gasification in the presence of a catalyst to form hydrocarbon vapors, which are then condensed to liquid fuels. This procedure for synthesizing hydrocarbons is based on the work of Fischer and Tropsch in Germany in the 1920s. Just prior to and during World War II, Germany produced oil and gasoline by this process at a maximum rate of only about 15,000 bbl/day because of the small output of the individual reactors compared to that obtainable at the time with direct liquefaction. Development of the Fischer-Tropsch process has been pursued in South Africa from 1955 to the time of writing and continues to be worked on. The Sasol plants in that country employ the largest coal gasification banks in the world and produce over 100 million m³/day (3500 million scf/day) of medium-CV gas. The plants produce about 100,000 bbl/day of motor fuels, employing individual reactors with capacities about 100 times greater than those of the original commercial units in Germany.

Oil shale production has also had a long history, with the earliest shale oil industry started in France in 1838, where

oil shale, which is a sedimentary rock containing an insoluble hydrocarbon, was crushed and distilled to make lamp fuel. Its operation was intermittent until the late 1950s, when it was terminated. In 1862, production of oil from shale was begun in Scotland, where it ran for about a hundred years. It reached its peak in 1913, with the production of about 6 thousand bbl/day of shale oil. Many countries have had shale oil retorting (distilling) facilities including the United States, which has particularly large reserves of oil shale. But as a result of the volatile economics of production, oil shale development has been turned on and off in the United States for more than a century. In 1991 there was only one major commercial retorting facility, that of the Unocal Corp. at Parachute Creek, Colorado. Constructed in the 1980s to produce 10,000 bbl/day of shale oil, by 1991 the plant was producing shale oil at a rate of 6000 to 7000 bbl/day when running, which was about two-thirds of the time.

Tar sands, also called oil sands, which are a mixture of sand grains, water, and a high-viscosity crude hydrocarbon called bitumen, are found in every continent. The most sizable reserves are found in Canada and in Venezuela. Between 1930 and 1960 commercial enterprises were formed and reformed with regularity to exploit the large Athabasca deposits in Alberta, Canada. In 1965 commercial production was begun in integrated surface plants that extracted the bitumen from the tar sands with hot water and upgraded it by distillation and hydrogen addition (hydrotreating). The upgrading procedure is much the same as that used in the refining of natural crude oil, and by this means a high-quality synthetic crude is produced. In 1990 Canadian commercial production from its two largest surface plants amounted to over 210,000 bbl/day of synthetic crude.

II. THERMAL CONVERSION PROCESSES

A. Pyrolysis

Pyrolysis refers to the decomposition of organic matter by heat in the absence of air. A common synonym for pyrolysis is devolatilization. Thermal decomposition and destructive distillation are frequently used to mean the same.

When coal, oil shale, or tar sands are pyrolyzed, hydrogen-rich volatile matter is distilled and a carbon-rich solid residue is left behind. The carbon and mineral matter remaining behind is the residual char. Pyrolysis is one method to produce liquid fuels from coal, and it is the principal method used to convert oil shale and tar sands to liquid fuels. Moreover, as gasification and liquefaction are carried out at elevated temperatures, pyrolysis may be considered the first stage in any conversion process.

The use of pyrolysis for the production of liquid products is illustrated in the block diagram in Fig. 1. The py-

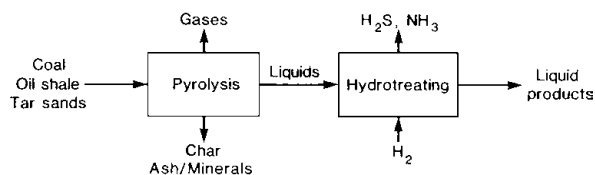


FIGURE 1 Pyrolysis. [Reprinted with permission from Probst, R. F., and Hicks, R. E. (1990). "Synthetic Fuels," pH Press, Cambridge, MA.]

rolysis vapors, consisting of condensable tar, oil, and water vapor, and noncondensable gases, consisting mainly of hydrogen (H_2), methane (CH_4), and oxides of carbon (CO , CO_2), are produced by heating of the raw material. The char, ash, and minerals left behind are rejected. The hydrocarbon vapors are treated with hydrogen to improve the liquid fuel quality and to remove the sulfur and nitrogen which came from the original raw material. The sulfur and nitrogen are removed as hydrogen sulfide (H_2S) and ammonia (NH_3) gases which form as a result of the hydrogenation.

The composition of the raw material is important in determining the yield of volatile matter, while the pyrolysis temperature affects both the amount and the composition of the volatile yields. When coal, oil shale, and tar sand bitumen are heated slowly, rapid evolution of volatile products begins at about 350 to $400^\circ C$, peaks sharply at about $450^\circ C$, and drops off very rapidly above $500^\circ C$. This is termed the stage of "active" thermal decomposition. There are three principal stages of pyrolysis. In the first stage, above $100^\circ C$ and below, say, $300^\circ C$, the evolution of volatile matter is not large and what is released is principally gas composed mainly of carbon dioxide (CO_2), carbon monoxide (CO), and water (H_2O). In the active or second stage of decomposition, about three-quarters of all the volatile matter ultimately released is evolved, with methane the principal noncondensable gas. The third stage is the one most appropriately defined for coal, in which there is a secondary degasification associated with the transformation of the char, accompanied by the further release of noncondensable gases, mainly hydrogen.

The total volatile matter yield, and hence the yield of tar plus light oils, is proportional to the hydrogen-to-carbon ratio in the raw material. On the other hand, the chemically formed water vapor that distills off during pyrolysis in an inert atmosphere is proportional to the oxygen-to-carbon ratio. The yields and product distributions also depend on the rate of pyrolysis.

B. Gasification

Gasification is the conversion of a solid or a liquid into a gas. In a broad sense it includes evaporation by heating, although the term is reserved for processes involving

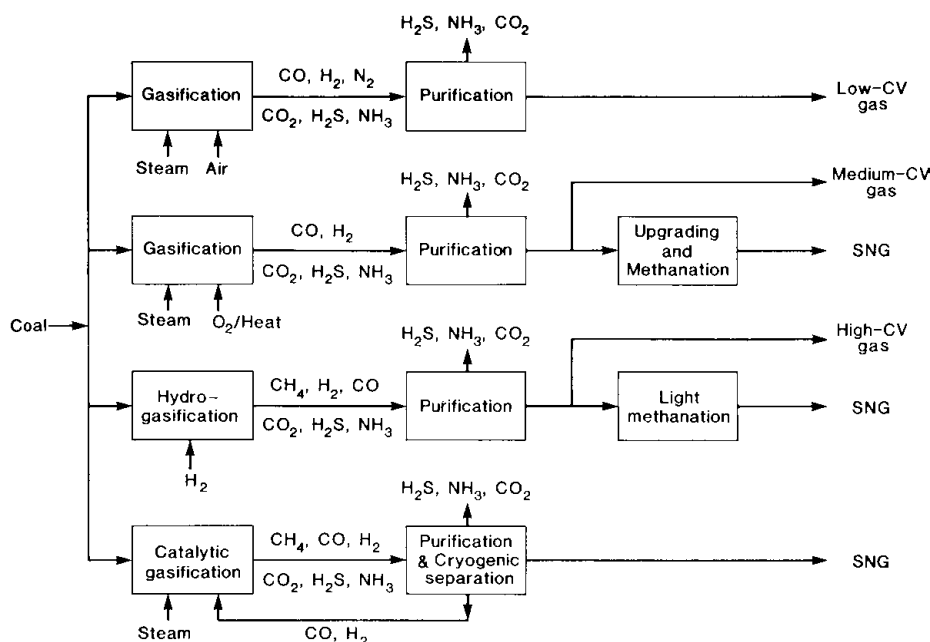


FIGURE 2 Gasification of coal. [Reprinted with permission from Probstein, R. F., and Hicks, R. E. (1990). "Synthetic Fuels," pH Press, Cambridge, MA.]

chemical change. The primary raw material for gasification is normally considered to be coal, although the use of oil shale for gasification has been discussed. Pyrolysis of coal is one method of producing synthetic gas and was the method pioneered in the early nineteenth century. Today the principal methods considered or in use for the gasification of coal to produce synthetic gases are shown in Fig. 2.

The most widely used technologies for the manufacture of gas employ indirect hydrogenation by reacting steam with coal in the presence of either air or oxygen. When air is used, the product gas will be diluted with nitrogen (N_2) and its calorific value will be low in comparison with that of the gas manufactured using oxygen (O_2). The dilution of the product gas with nitrogen can be avoided by supplying the heat needed for the gasification from a hot material that has been heated with air in a separate furnace or in the gasifier itself before gasification. In all of the cases, the gas must be cleaned prior to using it as a fuel. This purification step involves the removal of the hydrogen sulfide, ammonia, and carbon dioxide, which are products of the gasification. As with pyrolysis, the hydrogen sulfide and ammonia are formed from the hydrogenation of the sulfur and the nitrogen that were originally in the coal.

Medium-CV gas, consisting mainly of carbon monoxide and hydrogen, can be further upgraded by altering the carbon monoxide-to-hydrogen ratio catalytically and then, in another catalytic step, converting the resulting "synthesis" gas mixture to methane. A high-CV gas can

be produced by direct hydrogenation, termed hydrogasification, in which hydrogen is contacted with the coal. A procedure that allows the direct production of methane is catalytic gasification. In this method the catalyst accelerates the steam gasification of coal at relatively low temperatures and also catalyzes the upgrading and methanation reactions at the same low temperature in the same unit.

A simplified representation of steam-oxygen or steam-air gasification of coal is shown in Fig. 3. The gasifier represented is termed a moving bed gasifier, in that crushed coal enters the top of the gasifier and moves downward at the same time that it is being reacted, eventually being

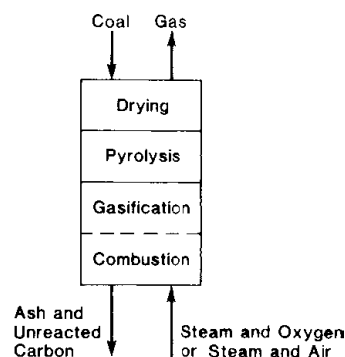


FIGURE 3 Schematic of a moving bed gasifier. [Reprinted with permission from Probstein, R. F., and Hicks, R. E. (1990). "Synthetic Fuels," pH Press, Cambridge, MA.]

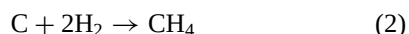
removed from the bottom as ash and any unreacted coal. The coal that enters the gasifier at the top is first dried by rising hot gases. Further heating results in devolatilization and pyrolysis. The next stage is the gasification zone, where the temperatures are typically above about 800°C and below about 1500°C. The temperatures are controlled by the temperatures in the combustion zone, which are a function primarily of the relative level of oxygen put into the gasifier.

The chemical reactions that take place in the gasifier are most easily presented by representing the coal by pure carbon (C). In the combustion zone the carbon is burned with the oxygen to produce carbon monoxide and carbon dioxide with the release of heat. The gasification chemistry is more complex but may be represented by a few principal reactions, as oxygen may be assumed not to be present beyond the combustion zone. In the gasification zone the carbon reacts with the steam put into the gasifier to produce carbon monoxide and hydrogen, using the heat released by the combustion; that is, it is an endothermic reaction. This endothermic reaction is known as hydrolysis, although it is more frequently referred to as the carbon–steam or gasification reaction. It may be written in chemical notation as



Atoms are conserved in a chemical reaction so that the number of carbon atoms, oxygen atoms, and hydrogen atoms must be the same on each side of the equation. The formula states that one atom of carbon reacts with one molecule of water to form one molecule of carbon monoxide and one molecule of hydrogen gas. The relative amounts of the substances participating in a reaction are given by the coefficients in the reaction formula, termed stoichiometric coefficients. In this case all the stoichiometric coefficients are one.

The carbon will also react with the hydrogen produced, to form methane. This reaction is termed hydrogenolysis or, more often, the carbon–hydrogen or hydrogenation reaction. It releases heat; that is, it is exothermic and may be written



In the oxygen-depleted gasification zone, the coal may also “burn” in the carbon dioxide and form carbon monoxide following the endothermic Boudouard reaction



Other reactions take place but are not discussed here. It is noted only that even at equilibrium the relative amounts of different gases produced will depend on the temperature and pressure in the gasifier and on the amount of steam

and oxygen relative to the amount of carbon put into the system.

C. Synthesis

The raw gas produced on gasification of coal has a low to medium calorific value, depending on whether air or oxygen is used as the oxidant in directly heated gasifiers. The product from indirectly heated gasifiers, in which an inert material is typically used to transfer the heat from an external source, is generally a medium-CV gas. One of the major reasons for producing synthetic fuels is to replenish dwindling natural supplies of traditional fuels such as natural gas and gasoline. A second reason is to eliminate pollutants to provide a clean-burning fuel. Removal of ammonia, hydrogen sulfide, and inert gases is an obvious requirement and has been noted. The principal gaseous products from gasifiers are carbon monoxide and hydrogen, which, although useful as a fuel, are not direct replacements for natural gas. These products can, however, be reacted with steam to produce substitute natural gas (SNG) as indicated by the methanation blocks in Fig. 2. These same products can also be reacted to produce gasoline, methanol, and other liquid fuels. Production of liquid fuels from coal after first completely breaking down the coal structure in a gasification step is known as “indirect liquefaction” and is shown schematically in Fig. 4.

A gas in which the major active components are carbon monoxide and hydrogen is called a synthesis gas, as these two compounds can be made to combine, or synthesize, to form a large number of products. The products formed from the synthesis gas depend both on the hydrogen-to-carbon monoxide ratio in the gas and on the catalyst and reactor conditions. Hydrogen-to-carbon monoxide mole ratios range from 3, for methane production with the rejection of water, to 1 to 0.5, for gasoline production with the rejection of carbon dioxide. The required hydrogen-to-carbon monoxide ratio can sometimes be achieved directly in the gasifier, although a H_2/CO ratio as high as 3 is normally not produced in commercial systems. In fact, many gasifiers produce a gas having a H_2/CO ratio of less than 1. In these cases an adjustment to the H_2/CO ratio is normally required, and is done by adding steam to the synthesis gas

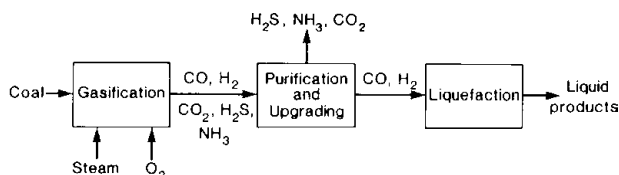
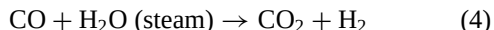


FIGURE 4 Indirect liquefaction of coal. [Reprinted with permission from Probst, R. F., and Hicks, R. E. (1990). “Synthetic Fuels,” pH Press, Cambridge, MA.]

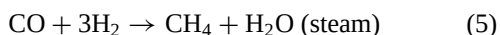
and reacting it with carbon monoxide to form hydrogen and carbon dioxide:



This is called the water–gas shift reaction or, frequently, just the shift reaction. The shift reaction is moderately exothermic; that is, it releases heat. Optimum operating temperatures are low, usually below about 225°C. The reaction will not proceed appreciably unless catalyzed, traditionally by reaction over an iron/chromium catalyst.

The need to “shift” the gas introduces an additional process step, so increasing overall the process complexity. In cases where the required synthesis gas composition can be achieved directly in the gasifier, this may be preferred in the interest of reducing the complexity.

Of interest in synthetic fuel manufacture is the production of SNG, which is principally methane. Methane (CH_4) does not contain oxygen, and the oxygen in the carbon monoxide may be rejected as either water or carbon dioxide. Typically water is rejected following the reaction



In this reaction a H_2/CO ratio of 3 is required in the synthesis gas, and one-third of the hydrogen content is wasted in rejected steam. This reaction is carried out over a zinc/chromium catalyst and is highly exothermic.

Perhaps the simplest synthesis reaction is the combination of one molecule of carbon monoxide with two molecules of hydrogen to form methanol (CH_3OH)



The catalyst used for this reaction is a copper-containing one, with reaction temperatures of about 260°C and pressures down to about 5 MPa. As with methane manufacture, the reaction is an exothermic one.

Finally, we note the commercially important Fischer–Tropsch synthesis reaction for gasoline manufacture, mentioned in Section I.B. The reaction formula may be written



Here the chemical formula is written CH_2 , which is one-eighth of a typical gasoline molecule (C_8H_{16}). The reaction is catalyzed by a number of metal-based catalysts including iron, cobalt, and nickel. The reactors in which the synthesis takes place operate within a temperature range of 225 to 365°C and at pressures from 0.5 to 4 MPa. It should also be noted that the Fischer–Tropsch reactions produce a wide spectrum of oxygenated compounds such as alcohols.

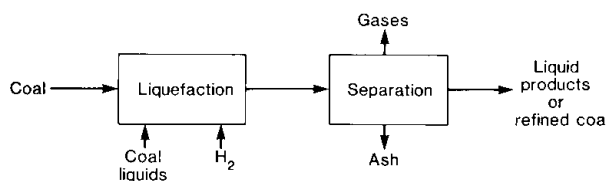
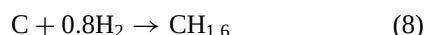


FIGURE 5 Direct liquefaction of coal. [Reprinted with permission from Probst, R. F., and Hicks, R. E. (1990). “Synthetic Fuels,” pH Press, Cambridge, MA.]

D. Direct Liquefaction

The two principal routes for the direct hydrogenation of coal to form a liquid involve the addition of hydrogen to the coal either directly from the gas phase or from a donor solvent. When the hydrogen is added directly from the gas phase, it is mixed together with a slurry of pulverized coal and recycled coal-derived liquid in the presence of suitable catalysts. This is called hydroliquefaction or catalytic liquefaction and is essentially the Bergius technology mentioned in Section I.B. In the donor solvent procedure a coal-derived liquid, which may or may not be separately hydrogenated, transfers the hydrogen to the coal without external catalyst addition. These procedures are illustrated schematically in the block diagram in Fig. 5.

The direct liquefaction of coal may be simplistically modeled by the chemical reaction



Direct liquefaction processes under development are typically carried out at temperatures from about 450 to 475°C and at high pressures from 10 to 20 MPa and up to 30 MPa. Despite the slow rate at which liquefaction proceeds, the process itself is thermally rather efficient, since it is only slightly exothermic. However, hydrogen must be supplied and its manufacture accounts for an important fraction of the process energy consumption and cost of producing the liquid fuel. The hydrogen itself may be produced, for example, by the gasification of coal, char, and residual oil.

III. TECHNOLOGIES

A. Gas from Coal

The three principal reactor types employed in coal gasifier design are the moving packed bed, the entrained flow, and the fluidized bed reactor. In the discussion of gasification principles the moving packed bed (Fig. 3) was used to illustrate steam–oxygen or steam–air gasification of coal.

The reactor type strongly influences the temperature distribution and, in this way, the gas and residue products. The reaction temperature typically varies from about 800 to 1500°C, and up to a maximum of about 1900°C in entrained flow oxygen reactors. Each type of gasifier

covers a specific temperature range. At high temperatures a synthesis gas is produced and at low temperatures methane formation is favored. Gasifiers in which the temperature is low enough that the residual ash does not melt are sometimes referred to as "dry ash gasifiers." High-temperature gasifiers in which molten ash (slag) is formed are called "slagging gasifiers." The slagging temperature is dependent on the ash composition but, for most coals, lies roughly in the range 1200 to 1800°C.

Moving bed coal gasifiers (see Fig. 3) operate with countercurrent flow and use either steam and oxygen or steam and air, and the residue may be either slag or dry ash plus any unconverted carbon. Coal particles in the size range of 3–50 mm are fed into the top of the gasifier. The coal passes downward, with an average linear bed velocities of the order of 0.5 m/hr in atmospheric steam/air gasifiers and 5 m/hr in high-pressure steam/oxygen gasifiers.

Representative of the moving bed gasifier are the Lurgi dry ash and slagging gasifiers. The Lurgi dry ash gasifier was the first high-pressure gasifier and was introduced in commercial operation in Germany in 1936. Nominal operating pressures of present commercial units are about 3 MPa, although they have been run at 5 MPa, with projected operating pressures up to 10 MPa. Temperatures in the combustion zone range from about 1000 to 1400°C, and those in the gasification zone from about 650 to 800°C. Typical coal throughputs are 800 t/day. The gasifiers are about 4 to 5 m in diameter and about three times as high, excluding the coal feed and ash lock hoppers that are attached to the top and bottom, respectively, and that more than double the height. A schematic drawing of the Lurgi dry ash gasifier is shown in Fig. 6.

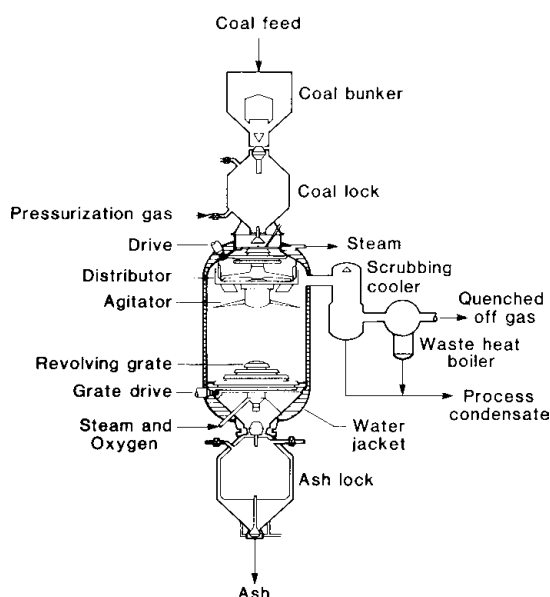


FIGURE 6 The Lurgi dry ash gasifier.

Not shown in Fig. 6 are the gas cleaning and purification units for the product gases leaving the gasifier. As discussed previously, in the manufacture of synthetic fuels any hydrogen sulfide, ammonia, and carbon dioxide present in the product (or byproduct) gases from a reactor usually must be removed. Ammonia is very soluble in water and is generally removed by washing the gases with water. Most of the hydrogen sulfide and carbon dioxide must be removed by other means. The procedure generally used is to remove the hydrogen sulfide and carbon dioxide, which are called acid gases, by absorption into an appropriate liquid solvent. The gases are subsequently desorbed from the liquid by heating and/or pressure reduction. The hydrogen sulfide is then converted to elemental sulfur, in part by burning it in a procedure known as the Claus process.

Entrained flow gasifiers all use coal (or char) pulverized to a size of the order of 75 μm . Oxygen or air together with steam generally is used to entrain the coal, which is injected through nozzles into the gasifier burner. Hot product gas may also be employed to entrain the coal and at the same time gasify it. In the Texaco gasifier, which is the gasifier in use at the Cool Water plant mentioned in Section I.B, the solids are carried in a water slurry, pumped up to gasification pressure (4 MPa), transported to the top of the gasifier, and injected through a burner into the reactor together with oxygen. The most important feature of entrained flow gasifiers is that they operate at the highest temperatures under conditions where the coal slags. In the Texaco gasifier, for example, the gasification temperature is about 1400°C.

Fluidized bed gasifiers are fed with pulverized or crushed coal that is lifted in the gasifier by feed and product gases. In single-stage, directly heated gasifiers a steam/oxygen or steam/air mixture is injected near the bottom of the reactor, either cocurrently or countercurrently to the flow of coal (or char). The rising gases react with the coal and, at the same time, maintain it in a fluidized state. Fluidization refers to the case in which the force of the gas on the particles lifts them so that they are in balance against their own weight. The particle "bed" is then expanded typically to twice its settled height, and is in a locally stable arrangement which resembles a boiling liquid. As the coal is gasified, the larger-size mineral particles, which are about twice as dense as the carbonaceous material, fall down through the fluidized bed together with the larger char particles. The advantage of this procedure is that it provides for good mixing and uniform temperatures in the reactor. Although fluidized bed gasifiers are thought of as a relatively recent development, work on the Winkler fluidized bed gasifier began in Germany in 1921, and the first commercial unit went into operation in 1926.

One gasification procedure that is markedly different, although the chemistry is not, is that of underground, or *in situ*, gasification. In this method, the gasification is carried out directly in the unmined coal deposit, which, by appropriate preparation, is turned into a fixed packed bed. The reactants are brought down to the coal bed and the gases formed are brought up to the surface through holes drilled into the deposit.

B. Liquids and Clean Solids from Coal

The three principal routes by which liquid fuels can be produced from coal have been noted to be pyrolysis, direct liquefaction, and indirect liquefaction. A clean fuel that is a solid at room temperature can also be produced by direct liquefaction processes.

In pyrolysis processes the main limitation is that the principal product is char, so that the effectiveness of any technology rests on the ability to utilize the char, for example, to produce gas or electricity. A wide number of technologies were under large-scale development through the early 1980s. One that attained commercial status was the Lurgi–Ruhrgas process, which feeds finely ground coal and hot product char to a chamber containing a variable-speed mixer with two parallel screws rotating in the same direction. Temperature equalization and devolatilization are very rapid due to the uniform mixing and high rates of heat transfer. The pyrolysis liquids and gases are removed overhead from the end of the chamber. Some of the new product char is burned in a transfer line and recycled to the reactor to provide the heat for the pyrolysis.

One process that was developed but not commercialized was the TOSCOAL process, in which crushed coal is fed to a horizontal rotating kiln. There it is heated by hot ceramic balls to between 425 and 540°C. The hydrocarbons, water vapor, and gases are drawn off, and the char is separated from the ceramic balls in a revolving drum with holes in it. The ceramic balls are reheated in a separate furnace by burning some of the product gas.

A process pioneered by the National Coal Board in England that has not reached the fully developed stage but that has considerable potential is supercritical gas extraction. In this process the coal is pyrolyzed at a relatively low temperature, around 400°C, in the presence of a compressed “supercritical gas,” that is, a gas whose temperature is above the critical temperature at which it can be liquefied. Suitable gases are, for example, a number of petroleum fractions. Under these conditions at high pressures, around 10 MPa, the gas density is like that of a liquid, and the gas acts like a strong solvent that causes the liquids to volatilize and be taken up by the vapor. By transferring the gas to a vessel at atmospheric pressure, the density of the solvent gas is reduced and the extracted

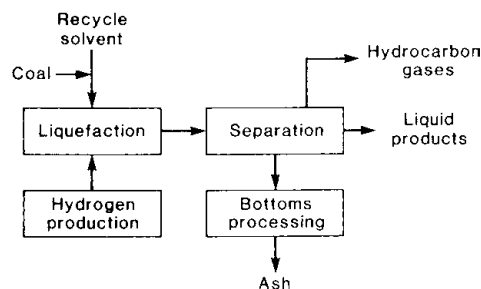


FIGURE 7 Generalized direct liquefaction process train. [Reprinted with permission from Probst, R. F., and Hicks, R. E. (1990). "Synthetic Fuels," pH Press, Cambridge, MA.]

tar precipitates out. The product is a low-melting glassy solid that is essentially free of mineral matter and solvent, and contains less nitrogen and sulfur than the coal.

Processwise two principal methods of direct liquefaction have been distinguished, in which the hydrogen may be added directly from the gas phase or a coal-derived liquid transfers the hydrogen to the coal. Despite the seeming difference, the major elements of both processes are similar as illustrated in the block diagram in Fig. 7. Coal is slurried with recycled oil or a coal-derived solvent, mixed with hydrogen, and liquefied at high pressures—in the case of hydroliquefaction, in the presence of an externally added catalyst. The resulting mixture is separated into gas and liquid products and a heavy “bottoms” slurry containing mineral matter and unconverted coal. Generally a large fraction of the carbon in the coal ends up in the bottoms, and most processes gasify this slurry to produce fuel gas and hydrogen.

Representative of the hydroliquefaction procedures in which hydrogen is added to the coal in the presence of a catalyst is the H-Caol process developed by Hydrocarbon Research, Inc. This procedure went through a pilot plant development capable of processing 530 t/day of dry coal to about 1350 bbl/day of low-sulfur fuel oil or 190 t/day of coal to a synthetic crude before operation was terminated. The difference in feed rates results from the fact that, to produce a synthetic crude, more hydrogen must be added, resulting in an increase in the residence time in the reactor and hence a decrease in the coal feed rate. In the process, coal crushed to less than 0.2 mm and dried is slurried with recycle oil at a ratio typically between 2 and 3 to 1, and then pumped to a pressure of around 21 to 24 MPa. Compressed hydrogen produced by gasification is added to the slurry and the mixture is preheated to 340 to 370°C. The mixture is passed upward into a reactor vessel operated at temperatures of about 450°C. The reactor contains an active, bubbly bed of catalyst, which is kept in a fluidized state by internally recycling slurry. The reactor is called an “ebullated bed” reactor because there is no locally stable

fluidized arrangement in it but, instead, a fluidized bed with an active and ebullient character.

Illustrative of a plant in which a coal-derived liquid or “donor” solvent transfers hydrogen to the coal is the Advanced Coal Liquefaction Research and Development Facility at Wilsonville, Alabama. The nominal coal feed of the pilot plant is 5.4 t/day, and in 1986 the facility was operational. Other plants in the United States have been run on a considerably larger scale but have been shut down. The plant was originally constructed to study the Solvent Refined Coal process for manufacturing a clean solid fuel in one stage. It then evolved to a facility to study two-stage liquefaction processes for making liquid fuels. The principal product from the single-stage process is an ash-free, low-sulfur, pitch-like extract that is a solid at room temperature. The product was formerly called “solvent refined coal,” or SRC, and is now called “thermal resid,” or TR.

In the process, coal dried and pulverized to less than 3 mm is mixed with recycle solvent at a mass ratio of about 1.5 solvent-to-coal. The slurry is pumped together with hydrogen at from 10 to 14 MPa and preheated to 400 to 450°C. It then enters the thermal liquefaction unit, which is a vertical tube in which the three phases flow cocurrently upward. The residence time in the unit is typically about 30 min, and under these conditions most of the carbonaceous material dissolves. The ash and undissolved coal are separated from the product liquid by a procedure developed by the Kerr–McGee Corp. termed “critical solvent deashing.” The principal is similar to that of supercritical gas extraction, discussed above, in that it employs the increased dissolving power of a solvent near its critical temperature and pressure. The solvent is mixed with the slurry and dissolves the product liquid. The solids settle out and the heavy product is subsequently recovered by decreasing the solvent density by heating. In the two-stage operation the product is upgraded by catalytic hydrogenation to light liquid hydrocarbons. The reactor employed for this is the ebullated bed H-Oil reactor developed by Hydrocarbon Research, Inc., which is similar to the H-Coal reactor.

A modification of the solvent extraction process that was investigated extensively in the 1980s is called *coprocessing*, in which the coal is processed together with a crude oil. The objective is to upgrade the oil and to simultaneously liquefy the coal. The fraction of coal in the feed may be less than 10%, in which case the major objective is to upgrade the oil using the coal as a catalyst, or more than 60%, in which case the process more closely resembles a solvent extraction process such as described above for the Wilsonville plant but without recycle of the solvent.

Coprocessing reactors are designed to operate at temperatures of 400 to 500°C and at pressures from 8 to 30 MPa; a catalyst may be added to increase yields. Feed

oils may be residues from petroleum refining processes or even from other synthetic fuel processes. Under the action of high pressure and temperature, the large oil molecules are ruptured to light products, and the sulfur atoms may be removed as hydrogen sulfide. The liquefaction of the coal molecules occurs by a process of extraction into the oil or, at higher temperatures, may involve thermal rupturing of the bonds. If hydrogen is added to the reactors, the latter route is termed hydrothermal processing.

Several processes have been investigated at the pilot and process development scale, and although some plans were made for commercial demonstration, there were no major developments anticipated in 1991.

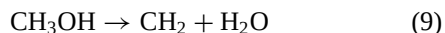
The last major category for the manufacture of liquid fuels is the indirect liquefaction procedures. The most extensive production of synthetic liquid fuels today is that being carried out by Fischer–Tropsch reactions at the South African Sasol complexes, with a combined output of over 100,000 bbl/day of motor fuels and other liquid products. The two largest plants, each with an output of about 50,000 bbl/day, employ 36 Lurgi dry ash, oxygen-blown gasifiers (see Fig. 6) apiece for the synthesis gas production. The gas is scrubbed with water for removal of particulate matter, tar, and ammonia, following which hydrogen sulfide and carbon dioxide are removed by absorption in cold methanol. The latter process is proprietary to Lurgi and is termed the Rectisol process.

The principal reactors used are fluidized bed reactors, called Synthol reactors, in which the feed gas entrains an iron catalyst powder in a circulating flow. The suspension enters the bottom of the fluidized bed reaction section, where the Fischer–Tropsch and the gas shift reactions proceed at a temperature of from 315 to 330°C. These reactions are highly exothermic, as described previously, and the large quantity of heat released must be removed. The products in gaseous form together with the catalyst are taken off from the top of the reactor. By decreasing the gas velocity in another section, the catalyst settles out and is returned for reuse. The product gases are then condensed to the liquid products.

Of the indirect liquefaction procedures, methanol synthesis is the most straightforward and well developed [Eq. (6)]. Most methanol plants use natural gas (methane) as the feedstock and obtain the synthesis gas by the steam “reforming” of methane in a reaction that is the reverse of the methanation reaction in Eq. (5). However, the synthesis gas can also be obtained by coal gasification, and this has been and is practiced. In one modern “low-pressure” procedure developed by Imperial Chemical Industries (ICI), the synthesis gas is compressed to a pressure of from 5 to 10 MPa and, after heating, fed to the top of a fixed bed reactor containing a copper/zinc catalyst. The reactor temperature is maintained at 250 to 270°C by injecting

part of the relatively cool feed gas into the reactor at various levels. The methanol vapors leaving the bottom of the reactor are condensed to a liquid.

An indirect liquefaction procedure of relatively recent origin is the Mobil M process for the conversion of methanol to gasoline following the reaction



The key to the process was the development by Mobil of a size-selective zeolite catalyst, whose geometry and pore dimensions have been tailored so that it selectively produces hydrocarbon molecules within a desired size range. This is a highly exothermic reaction and the major problem in any plant design is the reactor system to effect the necessary heat removal. A plant completed in 1985 in New Zealand uses about 4 million m³/day of natural gas as the feedstock to produce the methanol by the ICI procedure described above. In 1990 the plant produced about 16,000 bbl/day of gasoline, which is somewhat above its design output.

C. Liquids from Oil Shale and Tar Sands

Oil shale is a sedimentary rock containing the hydrocarbon “kerogen,” a high molecular mass organic material that is insoluble in all common organic solvents and is not a member of the petroleum family. Oil shale deposits occur throughout the world and may, in fact, represent the most abundant form of hydrocarbon on earth. The United States has by far the largest identified shale resource suitable for commercial exploitation in the Green River Formation in Colorado, Utah, and Wyoming.

Oil shale is characterized by its grade, that is, its oil yield, expressed as liters per ton (liters/t) in British units or as U.S. gallons per ton (gal/ton), as determined by a standard “Fischer” assay in which a given amount of crushed shale is pyrolyzed in a special vessel in the absence of air at 500°C. By definition, oil shale yields a minimum of 42 liters/t (10 gal/ton) of oil and may be found up to 420 liters/t (100 gal/ton). Lower-grade shale yields are below 100 liters/t. Commercially important western United States shales have an amount of organic matter (as mass% of the shale) of from 13.5 to 21%. In comparison, the organic matter in coal typically ranges from 75 to more than 90%, by mass. Consequently a significantly larger amount of oil shale must be processed compared to coal to obtain an equivalent hydrocarbon throughput. The inorganic content of oil shales is a mix of carbonates, silicates, and clays.

The principal method for producing oil from shale is by pyrolysis carried out in a vessel called a “retort,” with the process called “retorting” when applied to the commercial-scale recovery of shale oil. As with coal gasi-

fication, oil shale may be mined and retorted on the surface, or it may be retorted *in situ* and the released oil collected and pumped to the surface. Commercial-scale retorts are generally either moving packed beds or solids mixers. An *in situ* retort is in effect a moving bed reactor, but with the retorting zone moving through the stationary shale.

Oil shale retorts, like coal gasifiers, are classified according to whether they are directly or indirectly heated. In directly heated processes, heat is supplied by burning a fuel, which may be recycled retort off gas, with air (or oxygen) within the bed of shale. Some portion of either the coke residue or the unretorted organic matter may be burned as well. Not infrequently, most or even all the heat is provided by combustion of the kerogen. In indirectly heated processes a separate furnace is used to raise the temperature of a heat transfer medium, such as gas or some solid material such as ceramic, which is then injected into the retort to provide the heat. Whether the shale is heated by a gas or a solid defines two subclasses of the indirectly heated retort. The fuel that fires the furnace may be retort off gas or crude shale oil. The three heating methods for oil shale retorting are shown in Fig. 8.

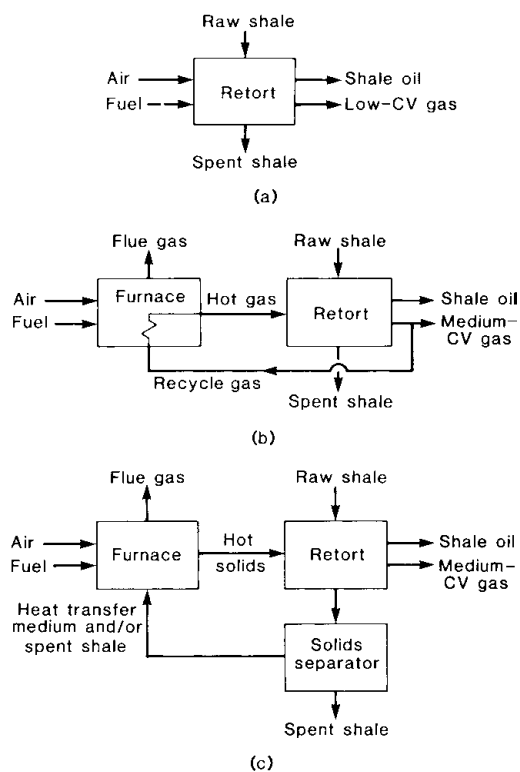


FIGURE 8 The three heating methods for oil shale retorting. (a) Directly heated retort; (b) indirectly heated retort, gas-to-solid heat exchange; (c) indirectly heated retort, solid-to-solid heat exchange. [Reprinted with permission from Probst, R. F., and Hicks, R. E. (1990). “Synthetic Fuels,” pH Press, Cambridge, MA.]

All surface processing operations involve mining, crushing, and then retorting. The liquid product of retorting is too high in nitrogen and sulfur to be used directly as a synthetic crude for refining and requires upgrading, for example, by treating with hydrogen, as discussed in connection with liquefaction, and/or by removal of carbon in a thermal distillation process termed coking. The spent shale remaining after retorting amounts to 80 to 85%, by mass, of the mined shale, so solid waste disposal is a major activity.

Most effort in the commercialization of oil shale processes has centered upon retort development. Over the years numerous technologies have been demonstrated for the surface retorting of oil shale, many of which have been discarded and then resurrected with modification, paralleling the on-again, off-again character of oil shale commercialization itself. Only a few will be mentioned here.

Two indirectly heated oil shale retorting technologies employing solid-to-solid heat transfer have been described in connection with coal pyrolysis. They are the TOSCOAL process, called the TOSCO process when used with oil shale, and the Lurgi-Ruhrgas process. The former process was fully developed before operations were terminated, and the latter has been commercialized in connection with coal devolatilization and hydrocarbon pyrolysis.

A retort that can be operated in either the direct or the indirect mode, which uses gas-to-solid heat transfer, is one developed by the Union Oil Co., now Unocal Corp. In this retort shale is charged into the lower and smaller end of a truncated cone and is pushed upward by a piston referred to as a "rock pump." In the indirect mode recycle gas that has been heated in a furnace flows in from the top countercurrent to the upward-moving shale. Combustion does not occur within the retort. As the shale moves upward it contacts the hot gas and is pyrolyzed. The shale oil flows down through the upward moving cooler fresh shale and is withdrawn from the bottom of the truncated cone together with the retort gas. In the directly heated mode, which has been demonstrated but discontinued, air without recycle gas is used, and nearly all of the energy of the residual carbon is recovered by combustion within the retort. The retort operating in the indirectly heated mode is the one being used in Unocal's 10,000 bbl/day facility at Parachute Creek, Colorado. Although toward the end of 1986 the plant was off-line for technical reasons, it was operating about one-half to two-thirds of the time between 1988 and 1991.

In situ retorting offers the possibility of eliminating the problems associated with the disposal of large quantities of spent shale that occur with surface retorting. *True in situ* (TIS) retorting involves fracturing the shale in place, ignit-

ing the shale at the top of the formation, and feeding in air to sustain the combustion for pyrolysis. The combustion zone moves downward, ahead of which is the retorting zone, and below that the vapor condensation zone. The gases and condensed oil and water are then pumped up from the bottom. Oil shale is not porous and generally does not lie in permeable formations, so adequate flow paths are difficult to create. To overcome this difficulty, an alternative approach known as *modified in situ* (MIS) has been developed. In this procedure a portion of the shale is mined out and the remaining shale is "rubblized" by exploding it into the mined void volume. The resulting oil shale rubble constitutes the retort.

Tar sands are normally a mixture of sand grains, water, and a high-viscosity crude hydrocarbon called bitumen. Unlike kerogen, bitumen is a member of the petroleum family and dissolves in organic solvents. At room temperatures the bitumen is semisolid and cannot be pumped, but at temperatures of about 150°C it will become a thick fluid. In the Alberta deposits of Canada, the bitumen is present in a porous sand matrix in a range up to about 18 mass%, although the sum of bitumen and water generally totals about 17%.

Two options for the recovery of oil from tar sands are of importance: mining of the tar sands, followed by above-ground bitumen extraction and upgrading; and *in situ* extraction, in which the bitumen is released underground by thermal and/or chemical means and then brought to the surface for processing or upgrading. Because the processes of *in situ* recovery are similar to those employed in the enhanced recovery of crude oil, they are not discussed.

Two surface extraction, full-scale commercial facilities are presently in operation to produce synthetic crude oil from the Alberta deposits. One, the Suncor, Ltd., facility was built in the late 1960s and, in 1991, was producing synthetic crude at a rate of about 58,000 bbl/day. The second and larger one, built in the mid to late 1970s, is the facility of Syncrude Canada, Ltd. In 1991 it was producing synthetic crude at a rate of about 156,000 bbl/day.

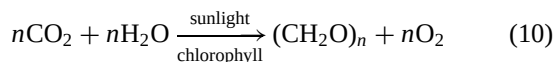
Both of the Canadian plants use the technique of hot water extraction to remove the bitumen from the tar sand. In this procedure the tar sand, steam, sodium hydroxide, and hot water are mixed and tumbled at a temperature of around 90°C. Layers of sand pull apart from the bitumen in this process. Additional hot water is added and the bitumen-sand mixture is separated into two fractions by gravity separation in cells in which the bitumen rises to the top and is skimmed off, while the sand settles to the bottom. The upgrading of the bitumen to a synthetic crude is then accomplished by oil refinery procedures including coking, in which carbon is removed by thermal distillation and hydrotreating.

IV. BIOMASS CONVERSION

A. Biomass as a Fuel Source

Biomass is any material that is directly or indirectly derived from plant life and that is renewable in time periods of less than about 100 years. More conventional energy resources such as oil and coal are also derived from plant life but are not considered renewable. Typical biomass resources are energy crops, farm and agricultural wastes, and municipal wastes. Animal wastes are also biomass materials in that they are derived, either directly or via the food chain, from plants that have been consumed as food.

As with conventional fuels, the energy in biomass is the chemical energy associated with the carbon and hydrogen atoms contained in oxidizable organic molecules. The source of the carbon and hydrogen is carbon dioxide and water. Both of these starting materials are in fact products of combustion, and not sources of energy in the conventional sense. The conversion by plants of carbon dioxide and water to a combustible organic form occurs by the process of photosynthesis. Two essential ingredients for the conversion process are solar energy and chlorophyll. The chlorophyll, present in the cells of green plants, absorbs solar energy and makes it available for the photosynthesis, which may be represented by the overall chemical reaction



$(\text{CH}_2\text{O})_n$ is used here to represent the class of organic compounds called carbohydrates or “hydrates of carbon,” several of which are made in the course of the reaction. Carbohydrates include both sugars and cellulose, which is the main constituent of the cell wall of land plants and the most abundant naturally occurring organic substance.

About one-quarter of the carbohydrate formed by photosynthesis is later oxidized in the reverse process of respiration to provide the energy for plant growth. The excess carbohydrate is stored. The plant typically contains between 0.1 and 3% of the original incident solar energy, which is a measure of the maximum energy recoverable from the plant if converted into a synthetic fuel. Some of this energy may, however, be degraded in the formation of intermediate products, and there will be additional losses in converting the biomass material into a conventional form.

One of the reasons for the great interest in biomass as a fuel source is that it does not affect atmospheric carbon dioxide concentrations. This is because carbon dioxide,

which is formed by respiration, biological degradation, or combustion, is eventually reconverted to oxidizable organic molecules by photosynthesis. Therefore, no net change in atmospheric carbon dioxide levels takes place provided an equivalent quantity of vegetation is replanted. More important, perhaps, is that this energy source is renewable. In addition, biomass fuels are clean-burning, in that sulfur and nitrogen concentrations are low, and because the hydrogen-to-carbon ratio is generally high. However, it is not expected that biomass will make a major contribution to overall energy requirements in the near future. The principal limitations of extensive biomass development are its high land and water requirements and the competition with food production.

B. Conversion Processes

The potential biomass conversion processes are shown in Fig. 9. They include biochemical conversion by fermentation and anaerobic digestion and the thermal processes of combustion, pyrolysis, and gasification. Fermentation produces mainly liquids, in particular, ethanol; pyrolysis results in both liquid and gaseous products; and gasification and anaerobic digestion produce gaseous fuels. Most biomass materials can be gasified, and the resulting gas may be used for synthesis of liquid fuels or substitute natural gas. Direct combustion of biomass is always an option and, in some instances, may be the only viable approach.

In principle, biomass resources can be converted using any of the biochemical or thermal conversion processes.

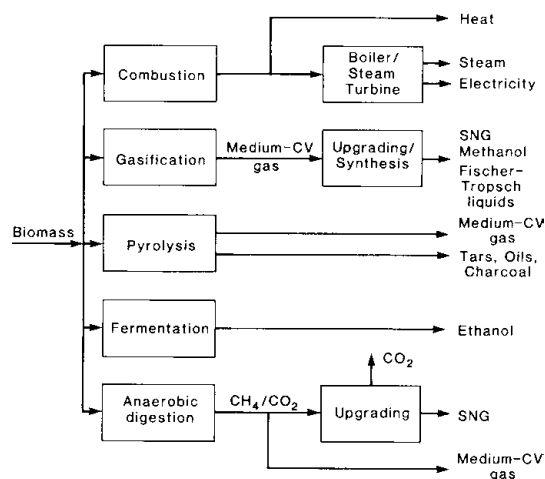


FIGURE 9 Biomass conversion processes. [Reprinted with permission from Probstein, R. F., and Hicks, R. E. (1990). "Synthetic Fuels," pH Press, Cambridge, MA.]

However, some processes can be expected to be more effective than others in recovering energy from specific resources. Wood is perhaps the most versatile resource, with the greatest potential. It is suitable for use on a large scale by combustion, or for air or oxygen gasification, and for pyrolysis. Municipal solid wastes, which are suitable for combustion and gasification on a large scale, also are considered to have potential. However, despite the many advantages of biomass, it is likely that only a small fraction of the world's energy needs could come from this source by the end of the twentieth century.

V. OUTLOOK

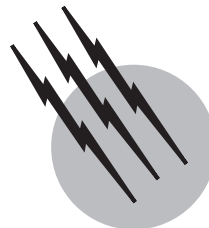
Most of the processes discussed either have been or are being used to supply synthetic fuels on a commercial basis. There is, therefore, little question as to the feasibility of these processes. In most cases, however, these ventures have proved and continue to prove economically unattractive in the face of abundant supplies of cheap natural gas and oil. When supplies dwindle and prices escalate, as is likely to happen eventually, specific processes can be expected to become marginally attractive. In the United States, probably the most competitive of the synthetic fuels are shale oil and low-CV and medium-CV gas. The more complex routes to liquid transportation fuels from coal can be expected to be more costly. In all cases a reduction in costs will occur as experience is gained from initial plants. Coal and, eventually, oil shale reserves will, however, also become depleted. Because biomass can probably make only a limited contribution to the total energy demand, other sources of energy will have to be harnessed. The development of synthetic fuels will probably be necessary to obtain the time needed for the evolution of such alternative energy sources.

SEE ALSO THE FOLLOWING ARTICLES

BIOENERGETICS • BIOMASS, BIOENGINEERING OF • BIOMASS UTILIZATION, LIMITS OF • BIOREACTORS • CATALYSIS, INDUSTRIAL • COAL STRUCTURE AND REACTIVITY • COMBUSTION • ENERGY EFFICIENCY COMPARISONS AMONG COUNTRIES • ENERGY FLOWS IN ECOLOGY AND IN THE ECONOMY • ENERGY RESOURCES AND RESERVES • RENEWABLE ENERGY FROM BIOMASS • WASTE-TO-ENERGY SYSTEMS

BIBLIOGRAPHY

- Beghi, G. E. (ed.) (1985). "Synthetic Fuels," D. Reidel, Hingham, MA.
- Elliott, M. A. (ed.) (1981). "Chemistry of Coal Utilization: Second Supplementary Volume," Wiley, New York.
- Gaur, S., and Reed, T. (1998). "Thermal Data for Natural and Synthetic Fuels," Marcel Dekker, New York.
- Klass, D. L. (1998). "Biomass for Renewable Energy, Fuels, and Chemicals," Academic Press, New York.
- Meyers, R. A. (ed.) (1984). "Handbook of Synfuels Technology," McGraw-Hill, New York.
- National Academy of Sciences (1980). "Energy in Transition 1985–2010," Final Report, Committee on Nuclear and Alternative Energy Systems, National Research Council, 1979, W. H. Freeman, San Francisco.
- Perry, R. H., and Green, D. W. (eds.) (1984). Fuels. In "Perry's Chemical Engineers' Handbook," 6th ed., pp. 9-3–9-36. McGraw-Hill, New York.
- Probstein, R. F., and Gold, H. (1978). "Water in Synthetic Fuel Production," MIT Press, Cambridge, MA.
- Probstein, R. F., and Hicks, R. E. (1990). "Synthetic Fuels," pH Press, MIT Branch P.O., Box 195, Cambridge, MA 02139. (First published 1982 by McGraw-Hill, New York.)
- Romey, I., Paul, P. F. M., and Imarisio, G. (eds.) (1987). "Synthetic Fuels From Coal. Status of the Technology," Graham and Trotman, Norwell, MA.
- Speight, J. G. (ed.) (1990). "Fuel Science and Technology Handbook," Marcel Dekker, New York.
- Supp, E. (1990). "How to Produce Methanol from Coal," Springer-Verlag, New York.



Thermal Cracking

B. L. Crynes

University of Oklahoma

Lyle F. Albright

Purdue University

Loo-Fung Tan

University of Oklahoma

- I. Introduction
- II. Major Feedstocks and Products
- III. Fundamental and Theoretical Considerations
- IV. Commercial Thermal Cracking
- V. Economics

GLOSSARY

Acetylenic Term describing hydrocarbons containing triple bonds, usually acetylene.

Acid gases Carbon dioxide and hydrogen sulfide, which are present in small quantities from the pyrolysis reactions.

Adiabatic Term describing an operation that occurs without the addition or removal of heat.

Endothermic reaction Reaction consuming heat as it proceeds.

Filamentous carbon Type of carbon that grows in long filaments or tubular structures on the inner walls of metal surfaces.

Hydrotreat To contact a hydrocarbon with hydrogen at moderate to high temperatures and pressures in order to perform hydrogenation reactions.

Pyrolysis gasoline Hydrocarbons formed during the

pyrolysis reactions that are within the gasoline range of boiling points.

Transfer-line exchanger (TLX or TLE) Primary heat exchanger adjacent to the pyrolysis furnace.

THERMAL CRACKING, or pyrolysis, is defined as the decomposition plus rearrangement reactions of hydrocarbon molecules at high temperatures. Hydrocarbons ranging from ethane, propane, n-butane, naphthas, and gas oils are used as feedstocks in pyrolysis processes to produce ethylene plus a wide variety of by-products, including propylene, butadiene, aromatic compounds, and hydrogen. Steam is, as a rule, mixed with the hydrocarbon feedstock. Thermal cracking is sometimes referred to as steam cracking, or just cracking. The emphasis in this article is on the production of ethylene and the above-mentioned by-products.

I. INTRODUCTION

A. Historical

Thermal cracking investigations date back more than 100 years, and pyrolysis has been practiced commercially with coal (for coke production) even longer. Ethylene and propylene are obtained primarily by pyrolysis of ethane and heavier hydrocarbons. Significant amounts of butadiene and BTXs (benzene, toluene, and xylenes) are also produced in this manner. In addition, the following are produced and can be recovered if economic conditions permit: acetylene, isoprene, styrene, and hydrogen.

Ethylene and propylene are used industrially in large quantities for the production of plastics and high molecular weight polymers and as feedstocks in numerous other petrochemical processes. Before the manufacture of ethylene from light paraffins (separated from natural gas) or petroleum fractions, ethylene was produced in the laboratory, and for commercial use, from fermentation-derived ethanol. It was also produced commercially from coke oven gas as early as 1920 and for several years thereafter. The technology developed in the processing of coal and the resulting coal-derived hydrocarbons was the foundation, to a considerable extent, of thermal cracking processes that have evolved for feedstocks obtained from petroleum and natural gas. With the development of ever-larger refining operations, numerous petrochemical developments followed. The discovery of plastics, such as polyethylene, polypropylene, and polystyrene, seeded the demand for ethylene, propylene, and aromatic compounds. Considerable research was conducted in the 1980s and the 1990s to develop improved methods of producing ethylene and other olefins. Methane (main constituent of natural gas), coal, methanol, garbage, wood, and shale liquids have, for example, been used as feedstocks. Such feedstocks have found no commercial applications. The current pyrolysis processes and feedstocks will almost certainly not be replaced in the foreseeable future.

Ethylene production has increased many fold in the last 40 to 50 years. In the United States, from 1960 to 2000, ethylene production increased from about 2.6 to 30 million metric tons/year while propylene production increased from 1.2 to 14 million tons/year. The growth rates on a yearly basis have, of course, depended in this time period on economic conditions in both the United States and worldwide. In 2000, worldwide production of ethylene was about 88 million tons/year; the production capacity was 104 million tons/year. In 1960, about 70% of both the ethylene and the propylene produced was in the United States. Relative growth rates in the last few years of both ethylene and propylene production have

been larger in Europe, Asia, and, more recently, the Near East. Currently, the United States produces only about 35% of the total ethylene and propylene. It should be emphasized that significant amounts of propylene are produced as a by-product in the catalytic cracking units of refineries. This propylene is sometimes separated and recovered as feedstocks to various petrochemical units. In the 1960s, studies were started relative to the interactions of reactor walls during pyrolysis reactions. More information on surface mechanisms follows later in this article.

II. MAJOR FEEDSTOCKS AND PRODUCTS

Feedstocks for various industrial pyrolysis units are natural gas liquids (ethane, propane, and n-butane) and heavier petroleum materials such as naphthas, gas oils, or even whole crude oils. In the United States, ethane and propane are the favored feedstocks due, in large part, to the availability of relatively cheap natural gas in Canada and the Arctic regions of North America; this natural gas contains significant amounts of ethane and propane. Europe has lesser amounts of ethane and propane; naphthas obtained from petroleum crude oil are favored in much of Europe. The prices of natural gas and crude oil influence the choice of the feedstock, operating conditions, and selection of a specific pyrolysis system.

Table I illustrates typical products obtained on pyrolyzing the relatively light feedstocks from ethane through butane, but significant variations occur because of the design and operating conditions employed with each light paraffin. The compositions of products obtained from naphthas, gas oils, and even heavier feedstocks differ to an even greater extent; the compositions of these heavier feeds vary over wide ranges. Tables II and III report typical

TABLE I Typical Primary Products from Light Feedstocks

Light feedstock	Product (wt%)			
	Ethane	Propane	n-Butane	i-Butane
H ₂	3.7	1.6	1.5	1.1
CH ₄	3.5	23.7	19.3	16.6
C ₂ H ₂	0.4	0.8	1.1	0.7
C ₂ H ₄	48.8	41.4	40.6	5.6
C ₂ H ₆	40.0	3.5	3.8	0.9
C ₃ H ₆	1.0	12.9	13.6	26.4
C ₃ H ₈	0.03	7.0	0.5	0.4
i-C ₄ H ₈				19.6
i-C ₄ H ₁₀	<0.2	<0.8	<1.9	20.0

TABLE II Typical Products from Heavy Feedstocks

Heavy feedstock	Product (wt%)		
	Naphtha	Gas oil	Vacuum distillate
CH ₄	10.3	8.0	6.6
C ₂ H ₄	25.8	19.5	19.4
C ₂ H ₆	3.3	3.3	2.8
C ₃ H ₆	16.0	14.0	13.9
C ₄ H ₆	4.5	4.5	5.0
C ₄ H ₈	7.9	6.4	7.0
BTX	10.0	10.7	18.9
C ₅ to 200°C (not BTX)	17.0	10.0	—
Fuel Oil	3.0	21.8	25.0
H ₂ + C ₂ H ₂ + C ₃ H ₄ + C ₃ H ₈	2.2	1.8	1.4

product mixtures for these heavy feedstocks. These three tables, along with other information, suggest the following general guidelines:

1. When ethane is the feedstock, the highest yields of ethylene are achieved, often as great as 80%. When propane and heavier feedstocks are used, yields are, however, less than 50%.
2. Propylene is the major olefin obtained during isobutane pyrolysis; however, there is no known industrial unit that uses it as the feedstock. Propylene yields are often in the 12–16% range when propane, heavier normal paraffins, naphthas, gas oils, and heavier petroleum feeds are pyrolyzed.
3. The heavier feedstocks produce appreciable amounts of butadiene; aromatic mixtures, commonly referred to as BTXs; and heavy nonaromatic compounds.
4. Coal oil and shale oils, containing appreciable amounts of aromatic compounds, result in correspondingly large amounts of BTXs.

TABLE III Typical Products from Nonconventional Feedstocks

Feedstock	Product		
	Coal Naphtha	Coal middistillate	Shale oil
H ₂	0.8	0.7	—
CH ₄	16	12	—
C ₂ H ₄	23	14	20–22
C ₃ H ₆	9	6	—
BTX	24	18	42–66
Fuel oil	24	47	—

III. FUNDAMENTAL AND THEORETICAL CONSIDERATIONS

A. Chemistry of Pyrolysis (Gas-Phase Reactions)

Understanding the mechanisms and kinetics of pyrolysis reactions has steadily advanced along with advances in technology. In the mid-1940s, free-radical reactions, as opposed to molecular schemes, were proposed to be the primary reaction steps. Pyrolysis reactions are usually divided into initiation, propagation and isomerization, and termination steps. Until relatively recently, these reactions were often thought to occur mainly, if not exclusively, in the gas phase. Table IV explains the main gas-phase reactions in a highly simplified manner for the pyrolysis of propane. Initiation steps, which are generally rate-controlling steps, are Reactions (1) and (2) of Table IV. In Reaction (1), the ethane molecule decomposes at the C–C bond to form two methyl radicals. These radicals react via Reaction (2) to form methane (an undesired by-product) and ethyl radicals. The propagation reactions, Reactions (3) and (4), occur relatively rapidly to produce ethylene and hydrogen (the two desired products). In theory, these two reactions continue until all of the propane has reacted to form the desired products. The number of times that they repeat themselves is referred to as the chain length. Eventually, there are termination steps resulting in a net destruction of free radicals. Reaction (5), which occurs in the gas phase, is just one termination step that can occur.

Termination reactions include other reactions in addition to Reaction (5) of Table IV. There are numerous free radicals present in addition to ethyl and methyl radicals. These other free radicals can also combine or couple. The coupling reactions in the gas phase, including Reaction (5), are highly exothermic. To promote such coupling in the gas phase, a relatively heavy molecule (or third body) is likely needed to help dissipate the exothermic heat of reaction.

Termination reactions also occur when a free radical in the gas phase reacts or couples with a free radical on a solid surface. The solid coke formed as a by-product during pyrolysis is essentially pure carbon, which has numerous free radicals on its surface. The exothermic heats of such

TABLE IV Simplified Ethane Pyrolysis Model

- (1) $\text{CH}_3\text{CH}_3 \rightarrow 2 \text{CH}_3^\bullet$
- (2) $\text{CH}_3^\bullet + \text{CH}_3\text{CH}_3 \rightarrow \text{CH}_4 + \text{CH}_3\text{CH}_2^\bullet$
- (3) $\text{CH}_3\text{CH}_2^\bullet \rightarrow \text{H}_2\text{C}=\text{CH}_2 + \text{H}^\bullet$
- (4) $\text{H}^\bullet + \text{CH}_3\text{CH}_3 \rightarrow \text{CH}_3\text{CH}_2^\bullet + \text{H}_2$
- (5) $\text{CH}_3^\bullet + \text{CH}_3\text{CH}_2^\bullet \rightarrow \text{CH}_3\text{CH}_2\text{CH}_3$

termination steps are dissipated into the coke. As will be discussed later, such surface termination steps are the start of a sequence of reactions that produce more coke.

B. Kinetics of Pyrolysis

Operating conditions, and particularly temperature, have a major effect on the kinetics of pyrolysis. Typical operating conditions are as follows:

Temperature of gaseous reactants, 750–1000°C
 Pressure, 2 to 6×10^5 Pa
 Residence time of reaction mixture in reactor coil,
 0.04–1.0 sec
 Steam/hydrocarbon weight ratio, 0.25–1.0

A plot of a laboratory reactor yields versus conversion of propane is shown in Fig. 1. These conversions were obtained over a range of 700–850°C. Conversions increase rapidly with temperature for all feeds at a given residence time. Conversions depend on temperature exponentially, and so even more rapid rates of reaction occur at higher temperatures.

Especially in the past, the kinetics of pyrolyses have often been reported as being proportional to the concentration of the feed hydrocarbon raised to some power. In such cases, the reaction order tends to shift to higher values with increased conversions and temperatures. These oversimplified models fail to account for the multiple reactions and products, but are reasonably successful for predicting the reactions, especially those at low conversions.

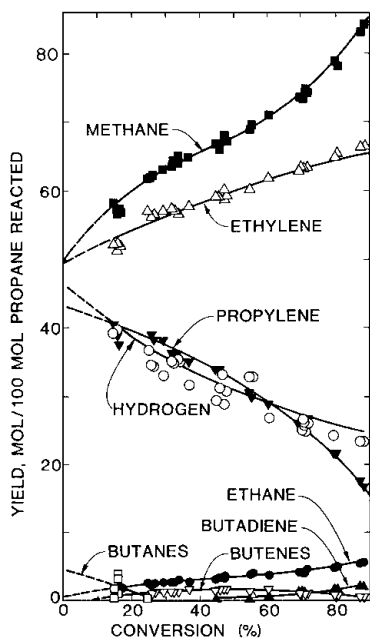


FIGURE 1 Propane conversion and yields.

Even though free-radical reaction schemes best represent the pyrolysis mechanisms, simple n th-order models for the pyrolysis of the lighter feedstocks have been proposed. The range of orders is from 0.5 to 2.0, and significant differences are often reported in the literature. The overall activation energies vary from about 50 to 95 kcal/g mol.

Pressure effects are, in general, not significant in the range of commercial pyrolysis interests. For hydrocarbon partial pressures of 0.2 to 2.0×10^5 Pa, few differences are seen. When pressures are increased to the range of 50 to 100×10^5 Pa, however, the global rate constants sometimes double.

In the past, reactor severity models were used for design purposes. These are models empirically relating some sort of reaction “severity factor” to temperature and time of reaction and/or concentrations. These can be thought of as a form of extent of conversion (severity) empirically related to reactor conditions. For example,

$$S = T \tau^a,$$

where S is the severity factor, T is the temperature, τ is the reactor residence time, and a is the empirical constant.

Steam is generally added for at least two reasons: first, it helps obtain quickly and then maintain the desired starting temperatures; second, it reduces the rate of coke collected on the inner surface of the reactor coil. Steam reacts slowly with deposited coke; nickel and iron on the coil surfaces, or present in the coke, catalyze this reaction.

The industrial furnaces are operated so that the temperature of the gaseous mixture increases steadily as it passes through the reactor coil. To obtain these increased temperatures, heat is transferred from the hot combustion gases surrounding the coil. In the inlet section of the coil, much of the heat transferred is employed to bring the gaseous mixture to the desired temperatures, and then pyrolysis starts at appreciable rates. In the latter portions of the coil, the heat transferred farther increases the temperature and, hence, increases the rates of reaction. Heat is, however, also needed to provide the large endothermic heats of reaction (for the initiation and propagation reactions, as shown in Table IV).

When lower temperatures are employed, longer residence times are required to obtain the desired conversions (or severity of pyrolysis). For the short residence time runs, in the 0.04- to 0.10-sec range, the rates of heat transfer are very large, requiring careful design of the equipment. Large temperature differences occur between the hot combustion gas surrounding the reactor coil and the reacting gases in the coil. There are substantial temperature differences across the walls of the coil and also between the inner surfaces of the coil and the reaction mixture. In those sections of the coil in which coke deposits on

the inner wall, there are also substantial temperature differences between the coke and the reacting gases. Temperatures of the coil surfaces and of the coke are unfortunately not well measured or known. Yet, reactions occur on these surfaces, as will be discussed in more detail later.

Much improved understanding of both the kinetics and the mechanisms in the pyrolysis of light paraffins has been realized using mathematical models containing terms for the numerous gas-phase reaction steps. Models containing hundreds or even reportedly over a thousand steps have been developed. Values of the activation energy and the frequency factor are used for each reaction step. Such parameters can generally be found in the literature or can be approximately based on literature results. Such models with the aid of powerful computers can be solved to predict conversions, yields, and product composition when temperature profiles are employed. Some models often approximate with considerable accuracy plant and pilot plant data. A few models have even incorporated terms for surface reactions. These surface reactions produce at least some of the coke, carbon monoxide, methane, and probably other compounds. A major problem in incorporating surface reactions into a model is that surface temperatures are higher, and not well known, as compared to the gas temperatures.

Considering the initiation reactions, such as Reactions (1) and (2) of Table IV, they are obviously of most relative importance when the concentrations of the hydrocarbon feedstocks are highest, i.e., during the initial phases of pyrolysis. The propagation reactions experience their most rapid kinetics when the concentrations of the free radicals and of the hydrocarbon feedstock are both relatively high, i.e., during the intermediate stages of pyrolysis. The termination reactions become increasingly important during the final stages of pyrolysis when the concentration of the hydrocarbon feed approaches low values. In commercial reactors, the highest temperatures occur during the final stages.

Mathematical models for the pyrolysis of naphthas, gas oils, etc. are relatively empirical. The detailed analysis of such a feedstock is essentially impossible, and all heavier feedstocks have a wide range of compositions. Such heavy hydrocarbons also contain a variety of atoms often including sulfur, nitrogen, oxygen, and even various metal atoms. Nevertheless, certain models predict the kinetics of pyrolysis, conversions, yields, etc. with reasonable accuracy and help interpret mechanistic features.

C. Surfaces Reactions in Pyrolysis Coils

Laboratory results in the last several years have demonstrated the importance of reactions occurring on the inner

surfaces of the pyrolysis coils and on the coke deposited on the coil surfaces.

1. Coke Formation

Coke, essentially pure carbon, is deposited on significant portions of the internal surfaces of all industrial coils. Surface reactions occur during the formation of this coke. The resulting coke is a highly undesired by-product for the following reasons:

1. The coke on the inner surface of the coil decreases heat transfer coefficients for the large amounts of heat that need to be transferred from the hot combustion gases in the furnace to the hydrocarbon-steam mixtures flowing through the coils. As a result of coke, smaller fractions of the heat of combustion are transferred to the reaction mixtures, and the already large costs of the fuel to the furnace are increased.
2. The coke deposited on the inner walls of the coil, and especially on the exit portion of the coil, increases the pressure drop of reaction gases as they flow through the coil. Hence, increased costs result, since the product gas mixture needs to be compressed as it enters the recovery and separation portion of the pyrolysis process.
3. At intervals often varying from 1 week to several months, a pyrolysis unit must be shut down in order to clean (or decoke) the coils. Such decokings sometimes require 1–2 days to complete. Obviously, as the rates of coking increase, the number of decokings per year also increases, causing the annual production rate of ethylene to decrease. Furthermore, utility and labor costs for each decoking are relatively expensive.
4. As will be discussed in more detail later, current methods of decoking contribute to decreased longevity of the metal coils. Further coking rates immediately after decoking are high for perhaps 1 day.
5. Ethylene yields are reduced. Laboratory tests at Purdue University when ethane was pyrolyzed indicate the coke formation is directly related to lower ethylene yields.

The following by-products are excellent coke precursors: acetylene, butadiene and other conjugated dienes, and aromatics. These compounds are produced mainly in the latter stages of pyrolysis. The following chemical sequences produce coke:

Mechanism 1 results in the formation of filamentous coke which has the following characteristics: is strongly

attached to the metal surface of the coil; initially, has small diameters but high length; and on occasion, the filaments are hollow or tubular. In the first reaction step of formation, nickel or iron carbides are formed from a precursor such as acetylene. The carbide particles are separated from the surface as coke formation occurs at the base of the particle. Generally, the particle is incorporated at the top of the filament. Mechanism 1 obviously corrodes the coil.

Mechanism 2 results in the production initially of tar droplets suspended in the gas phase. The exact sequence of reactions differs depending on the hydrocarbon feedstock. For ethane and propane feeds, the sequence is as follows: acetylene and/or dienes combine to form simple aromatics; these aromatics react to produce polyaromatics and eventually low-viscosity tars; next, agglomeration steps form droplets suspended in the high-velocity gas stream; the low-viscosity droplets dehydrogenate because of the high temperatures to produce high-viscosity droplets and hydrogen. These droplets collide with and often collect on the solid surfaces (either the inner surface of the coil or on the coke that has already been formed or collected). Rough surfaces and particularly filamentous coke on the surface are excellent collection sites. If a low-viscosity tar droplet hits the surface, the droplets spread out on the surface. With higher viscosities, the droplets often collect on the surface and retain their spherical shapes; electron microscope photographs often show collections of droplets looking somewhat like clusters of grapes. In other cases, the high-viscosity droplets sometimes rebound from the surface like a ping-pong ball. Laboratory studies indicate that gravity affects the rate of collection of the droplets; more droplets collect on top surfaces as compared to bottom surfaces. After the droplets collect on a hot surface, all carbon-hydrogen bonds break within a few minutes to form carbon and hydrogen. As C-H bonds break, free radicals form on the surface of the coke.

Mechanism 3 has, as a first step, reactions between surface radicals on the coke and the acetylene, butadiene, and gaseous free radicals; reactions probably also occur with ethylene and propylene. Reactions with gaseous free radicals were discussed earlier as a termination step in the gas-phase reactions. When acetylene reacts with the surface radicals, aromatic structures are formed on the surface. When the C-H bonds on the surface later break, graphitic coke is formed. The cokes produced by both Mechanisms 1 and 3 tends to be highly graphitic. Microscopic photographs have shown that Mechanism 3 thickens filamentous coke and causes spherical coke particles formed by Mechanism 2 to grow in diameter.

When coils produced of high-alloy steels are used, the initial layer of coke contains an appreciable fraction of filamentous coke, as indicated by microscopic

photographs of industrial coke samples. This coke is often rather porous with gas spaces around the filaments; porous coke obviously has relatively high resistances to heat transfer. Appreciable amounts of iron, nickel, and especially chromium are present in the initial layers of coke. The total amount of metals is probably in the 1–2% range. Coking rates are much higher while this coke is being formed as compared to coke produced later.

During latter stages of coking, the coke formed is generally solid and contains mainly cokes formed from Mechanisms 2 and 3. It contains less metal, perhaps in the 0.3–0.5% range; nickel, chromium, and iron are still present. On occasion, a metal fragment can be detected by the microscope with filamentous coke connected to it, i.e., a porous section of coke surrounded by solid coke.

Numerous laboratory and plant tests have been made in coils constructed of materials with nickel- or iron-free surfaces; e.g., quartz glass, silicon-coated and aluminum-coated steels, or ceramics. Filamentous coke was often completely absent, and the overall levels of coking were much reduced. Furthermore, the coking rates were essentially identical during the entire pyrolysis run.

2. Oxide Layer in Inner Surface of Coils in Furnaces

Extensive laboratory and plant data indicate that oxide layers form on the inner surfaces of high-alloy steel coils within several hours. Metal oxides and hydrogen form when metal atoms react with steam. These oxide layers may, with time, become as thick as 10–15 μ ; layers of 1–5 μ often form in 1–8 h. These oxide layers, however, have a much different metal composition as compared to that of the starting steel. For a stainless steel with a composition by weight of about 45–48% iron, 30% nickel, 20% chromium, and 0.3–0.7% manganese plus trace amounts of aluminum, titanium, niobium, carbon, etc., the surface composition, as measured by EDAX, may become approximately 65–75% chromium, 15–20% manganese, 5–8% iron, and 1–3% nickel. Occasionally, high levels of aluminum and titanium also form on the surfaces. Just below this oxide layer, a layer enriched with iron and nickel is formed. Clearly, there is a net diffusion of chromium, manganese, and sometimes aluminum and titanium toward the surface. Here, they are oxidized and form stable oxides. Because of the size and weight of the oxide molecules, further diffusion is essentially stopped. Iron and nickel form less stable oxides. The iron and nickel atoms tend to diffuse inward, forming an iron- and nickel-enriched layer below the oxide layer. Extensive experimental work performed at Purdue University clarifies the formation of these oxide layers on the surface of high-alloy steels. Gases with limited oxidative capabilities were

tested at 500–1000°C. 50:1 mixtures of hydrogen:steam and of CO:steam were found to be effective gases, along with pure CO, for pretreating the steels and forming oxide layers. Incoloy 800, HK-40, and HP steels were all investigated. As measured primarily by EDAX analysis and to a limited extent by ESCA analysis, significant changes in the surface composition of Incoloy 800 occurred with as little as 0.25 h pretreatment with one of the above three gases at 800–1000°C. After about 8 h of pretreatment, the chromium content at the surface had increased in some cases by at least 3 times, and the manganese content had increased by as much as 50 times. Material balances indicate that some manganese diffused by as much as 50–100 μ to reach the surface oxide layer. The rates of diffusion in HK-40 and HP steels were, however, much slower than those in Incoloy 800.

Purdue University investigations also found that hydrogen-steam mixtures slowly gasify (or remove) coke deposited on high-alloy steel coupons at 800–1000°C, i.e., the coupons are decoked. Following such a decoking, the coupon was exposed to pyrolysis conditions and decoked. The resulting coke adhered poorly to the coupon. A similar surface was also obtained when Incoloy 800 was pretreated with CO. Assuming such surfaces were formed in industrial coils, coke formation or collection might be completely eliminated. Because of the high velocities of the gases in the coil, any coke deposited would probably be quickly eroded and removed from the surface. For high-alloy steels, the oxide layers, whether on coils or coupons, have physical properties very different than those of the base metal (or starting steel). First, the surface layer minimizes the formation of filamentous coke because of the depletion of nickel and iron at the surface. Second, and unfortunately, the surface layer is relatively brittle and rather easily spalled off (or lost). The surface layer tends to crack due to the creep in the coils plus thermal expansion and contraction. Relatively large temperature changes occur for the surface layers when decoking is started and when it is completed. Vibrations also often occur while a pyrolysis plant operates. When spalling occurs, relatively large portions of the oxide layer are lost from the surface, hence exposing the iron- and nickel-rich layer that promotes filamentous coke production. Typical decoking procedures contribute substantially to both spalling and corrosion problems and to the high rates of coke formation once pyrolysis operations resume. During decoking, oxygen is normally employed, which raises the level of metal oxides on the surface. A net diffusion of iron and nickel atoms toward or to the surface likely occurs during decoking as a result. The metal-containing particles that spall off the surface during decoking may accumulate in the bottom sections of the coil and are gradually fluidized during subsequent pyrolysis operations. Within

several days of operation, some of the particles are trapped in the coke deposits.

There is no known experimental data on the fate of metal in the coke when the coke is gasified (forming carbon oxides) during decoking. Some of this metal may remain on the coil until pyrolysis operations are resumed.

In many pyrolysis units using ethane and/or propane as feeds, sulfur-containing additives are continuously added to the feed in small amounts. Typical additives include alkyl sulfides or disulfides, mercaptans, and hydrogen sulfide (H_2S). In the reactor coils, these compounds decompose, and at least some elemental sulfur is formed. Pretreatment of coils following decoking with desired amounts of these additives also reduces the level of coking when the pyrolysis operation is resumed. Such a pretreatment of the coil converts many metal oxides on the coil surface to metal sulfides; as a result, less coke is formed. It is highly important to use only optimum amounts of the additive. Too much additive results in increased amounts of coke and in excessive corrosion of the coils. Some nickel sulfides are liquids at coil temperatures and would be entrained in the gas stream. Optimum concentrations of the additive depend on the hydrocarbon used, operating conditions, and furnace used. Such additives are generally not used with naphtha or gas oil feedstocks, which generally contain sulfur compounds (often too many). Although ethylene producers hope their coils have service lives of 5–6 years, coils frequently need to be replaced sooner. That is not surprising since metal temperatures have tended to increase in the last 10 years in order to realize higher yields of ethylene. Metal temperatures in portions of the coil are currently often in the 1000–1075°C range. All of the following contribute to shorter coil life:

1. Creep in which vertical coils lengthen often by as much as several centimeters per year
2. Decreased wall thickness due to filamentous coke formation and spalling of metal oxide surfaces
3. Decreased chromium content of the high-alloy steels due to spalling and corrosion

High rates of coke formation plus frequent decokings accentuate all of the above problems. The following surface reactions occur during pyrolysis or decoking operations: oxidations to form metal oxides, reduction or partial reduction of oxides when they are contacted by the hydrocarbons in the gas phase, conversion of metal oxides to metal sulfides when sulfur-containing gases are present, conversion of metal sulfides to metal oxides due to reactions with steam or oxygen (during decoking). Such a complicated set of surface reactions promotes loss of metal atoms or compounds from the surface of the coil.

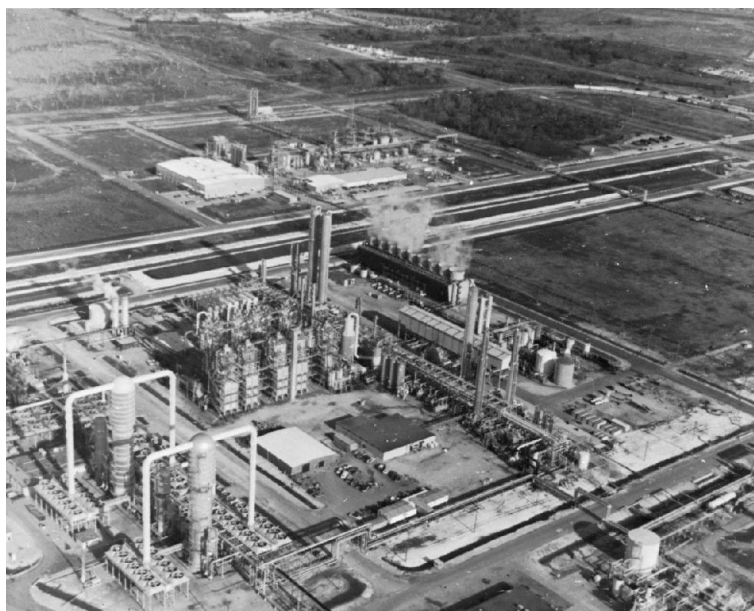


FIGURE 2 Dow Chemical ethylene plant. (Courtesy of Dow Chemical Company.)

IV. COMMERCIAL THERMAL CRACKING

A. General Process

Figure 2 is a photograph of a large, complex plant. Plants such as this contain reactor furnaces, distillation towers, heat exchangers, separators, dryers, compressors, and various other units as required by the specific feedstock and product distribution achieved. Figure 3 shows a simplified diagram of a plant.

1. Cracking Furnace and Reactor

For the primary step in the overall pyrolysis plant, the feedstock must be vaporized, if in liquid form, then mixed with steam, and finally preheated to the reactor temperature. When the feedstock (e.g., ethane or propane) is in gaseous form, vaporization is usually achieved by simple heat exchange with other product components such as condensing propylene. To vaporize liquid stocks such as naphthas, higher temperatures must be used. These feeds may be partially preheated before entering the furnace itself and then fully vaporized as they flow through the convective zone of the furnace. Typical furnace and tube geometries are shown in Fig. 4.

Typically, the tubing in the convection zone is placed in a “hairpin” fashion or series of connected U bends and is suspended horizontally. The steam that will be used as a diluent is also heated in the convection zone in other tubes. This superheated steam and the gaseous hydrocarbons are

combined and further heated to $\sim 700^{\circ}\text{C}$ by the end of the convection section. This mixture then passes to the radiant zone, where the main pyrolysis reactions occur and where the temperature is increased to around 1000°C .

Temperature is the most important operating variable in pyrolysis reactions; the large amounts of heat must be

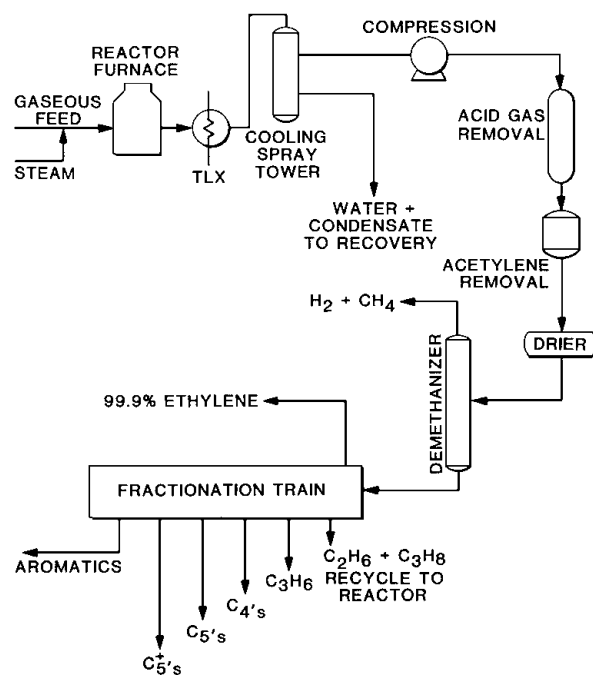


FIGURE 3 Simplified flow sheet of a general pyrolysis plant.

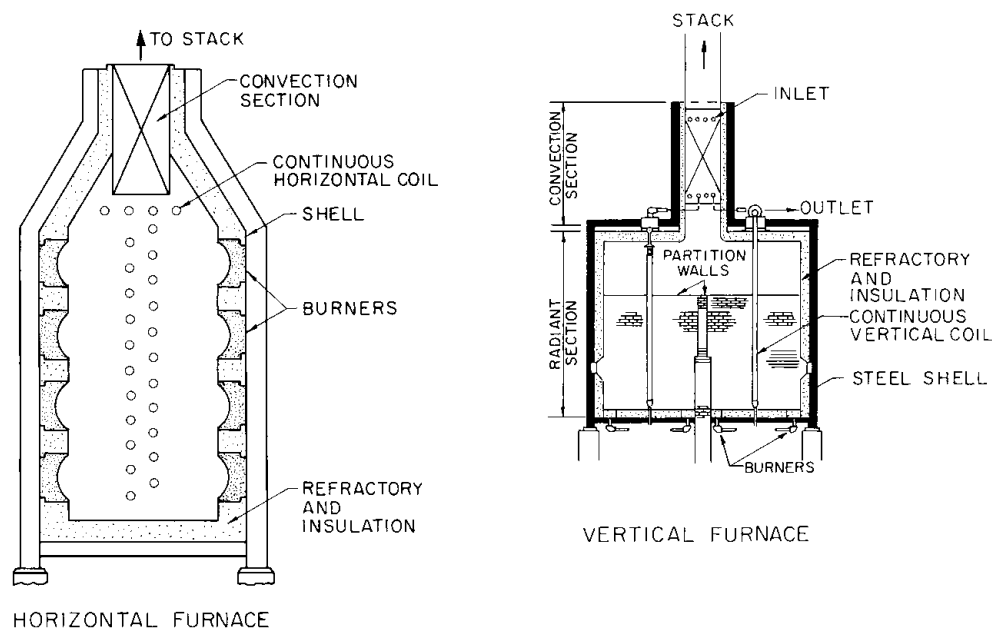


FIGURE 4 Furnace tube geometries.

transferred because most reactions are highly endothermic. As a general rule, higher temperatures (and the resulting lower residence times needed) are preferred for improved yields of ethylene, which is almost always the desired primary product. Hence, the reactor is normally fired as closely as possible to the limiting temperature of the reaction tubes.

Other variables of importance in designing these tubular pyrolysis reactors include the mass velocity (or flow velocity) of the gaseous reaction mixture in the tubes, pressure, steam-to-hydrocarbon-feedstock ratio, heat flux through the tube wall, and tube configuration and spacing. Pressure drop in the reactor is of major importance, especially because of the extremely high flow velocities normally employed.

The addition of steam to the entering feed provides several advantages, including heat sink which helps maintain higher temperatures and results in lower partial pressure of the hydrocarbons. The decreased partial pressure helps to minimize undesirable reactions. Increased amounts of steam result in increased steam-coke reactions, which result in the production of carbon oxides and in slower rates of coke formation on the metal surfaces. Typical steam requirements are shown in Table V.

High gas velocities (or high mass velocities) in the tubular reactor affect heat transfer through the boundary layer. There can be a rather large difference between the nominal process temperatures and tube skin temperatures. As already stated, there is a desire to take advantage of the tube material and to operate near its limit in order to re-

duce the actual residence time. As mass velocities increase within the reactor, higher heat transfer coefficients result. A countering effect of higher velocities, however, is increased pressure drop in the gas stream as it moves through the tube with a resultant increase in erosion at tube bends. There is a limit to the total pressure drop that can be tolerated with any given pyrolysis system from reactor tube inlet to outlet.

The reaction tubes (or coils, as they are often called) are positioned vertically in modern furnaces, but they are positioned horizontally in some furnaces of older design.

The furnaces are designed so that combustion of the fuel gas or liquid occurs around the reaction tubes. Temperatures of the combustion gases are often as high as 1200°C in the immediate vicinity of the tubes. Most heat is transferred by radiation, and the portion of the furnace in which the reaction tubes are located is often referred to as the radiant zone. The maximum permissible temperature for the metal in the tubes depends on the type of

TABLE V Typical Amounts of Steam for Commercial Pyrolysis

Feedstock	Weight ratio, steam/feed
Ethane	0.25–0.40
Propane	0.25–0.5
Naphthas	0.5–0.7
Gas oils	0.7–1.0

stainless steel, but is generally at most $\sim 1100^{\circ}\text{C}$; however, metal temperatures in the convection zone are significantly lower. The effluent combustion gases from the radiant zone exchange their heat in the convective section to preheat the feedstock and generate steam, as mentioned earlier.

Major progress has been made in the last 20 years, and especially in the late 1990s, in developing coated coils, which prevent the formation of filamentous coke. The coils of 10–20 years ago were coated with surfaces highly enriched with aluminum or coated with a thin layer of silica. Much reduced levels of coke formed on these coils, and decokings were, in general, much faster and easier to accomplish. Unfortunately, the coatings were lost (or destroyed) rather quickly, especially if higher temperatures were used. Examples of limited success have occurred, but there was clearly a need for improvement.

In the late 1990s, Surface Engineering Products and Alon Surface Technologies each developed coating procedures for high-alloy steels, such as HK, HP, Alloy 803, etc., that are used in ethylene coils. As of 2001, such coated coils have been installed in about 30 furnaces worldwide. These coils have demonstrated in most, if not all, cases much improved operation, including much reduced levels of coke formation, fewer decokings, easier and faster decokings, reduced CO formation, and increased annual levels of ethylene production. Such improvements are directly related to the fact that filamentous coke formation is much reduced, if not eliminated. Furthermore, the adhesion between the coke and the coil surface appears to be rather low. Frequent tests on the coils have indicated that most coatings are still in good condition after extended times of operation. Most coils will likely have lives of 5 or more years, except for several coils operated at relatively high temperatures.

Considerable fuel is required to provide the high temperatures necessary for pyrolysis of large plants. Attention to burner design and furnace details, as well as the choice of an economic fuel of reliable supply and composition, are, of course, mandated. The recovery and use of energy are dictated by economics. The flue gases from combustion are vented through stacks with appropriate attention to air pollution caused by incomplete combustion, stack visibility, and nitrogen oxide emissions. In general, gaseous fuels are preferred over liquid forms because of the ease of flow control, more complete combustion, and higher flame temperatures.

Tubular reactors normally take one of three geometries. The most common consists of tubing of the same diameter connected by U bends. The internal diameters are often 4 in. The total length of one reactor coil may be 300 ft. Four to eight coils are commonly placed within a single furnace.

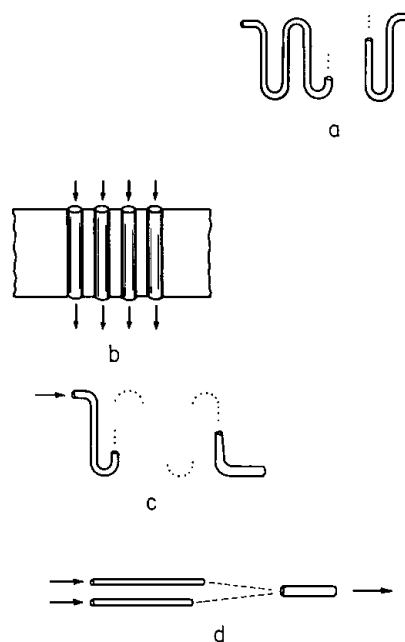


FIGURE 5 Tube designs: (a) same diameter, hairpins, which are the most common; (b) small diameter, single pass; (c) increasing diameter; (d) split coils; two small-diameter coils to one large.

Another configuration consists of the expanded and split coils that incorporate coil sections having two to three different diameters. Smaller diameter sections are connected (welded) by U bends to larger diameter sections. Often, two or three small-diameter sections arranged in parallel are connected to a single larger diameter tube (Fig. 5). Usually, the sections are positioned vertically in the furnace. The advantages of flexible residence times, reduction of pressure drops, and control of temperature in this more complex tube design are obvious. In a more recent design system, about 30–40 parallel and vertical tubes are fitted into a furnace; in this arrangement there are no U bends or any change of direction. The hydrocarbon-steam mixture simply passes through the radiant zone of the furnace and then exits into the transfer-line exchanger. These tube lengths are relatively short, up to 40 ft, and are of small diameter, up to 2 in. These furnace designs provide for extremely short residence times, less than 0.05 sec.

As indicated earlier, these configurations, almost without exception, are positioned vertically as they are suspended in the radiant zone of the furnace. Horizontally suspended tubes are more subject to buckling and sagging, especially at higher temperatures.

Multiple pyrolysis furnaces are employed in an industrial pyrolysis plant in order to maintain reasonably constant production levels, even when one furnace is shut down for decoking or maintenance repairs. The coils or tubes in a furnace or the transfer-line exchanger must

TABLE VI General Pyrolysis Tube Conditions

Condition	Value
Tube temperature	Up to 1050°C (1922°F)
Reaction gas temperature	Up to 950°C (1742°F)
Tube inlet total pressure	$3.7\text{--}6.4 \times 10^5$ Pa
Tube outlet pressure	$1.7\text{--}2.4 \times 10^5$ Pa
Heat flux	20,000–30,000 Btu/h $\times$ ft ²
Residence times	0.15–0.5 sec
Mass velocities	Up to 30 lb/ft ² $\times$ sec
Linear velocities	Up to 1000 ft/sec

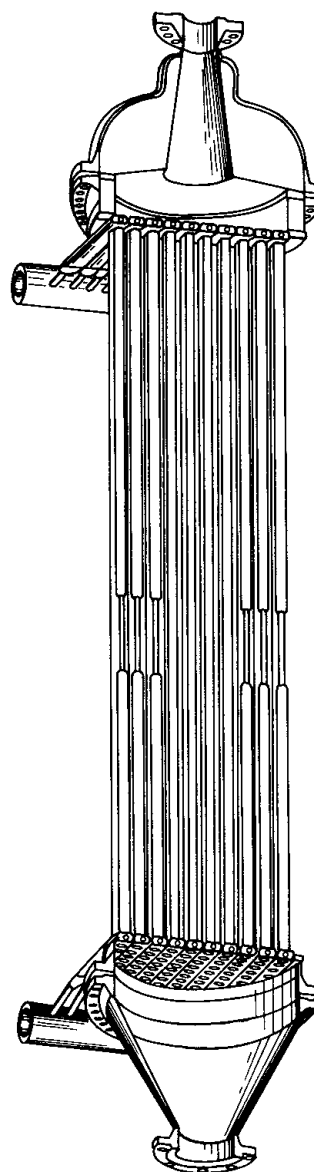
normally be decoked every few months. Hence, decokings are arranged as much as possible to be staggered between the various furnaces to maintain reasonably constant production rates and to utilize the operational and maintenance personnel effectively. Typical operating conditions for a pyrolysis furnace are summarized in Table VI.

2. Transfer-Line Exchangers

Transfer-line exchangers (frequently referred to as TLXs or TLEs) recover most of the sensible heat in the product gas stream exiting from the pyrolysis furnaces. These exchangers are located as close as possible to the furnaces in order to cool the product stream quickly and to maximize heat recovery. Sometimes as much as 5–10% of the reaction can occur in the transfer line itself, which is the unheated section of tubing outside the furnace and just ahead of the TLXs. The temperature of the product stream leaving the furnace may often be 800–850°C, depending to some extent on the feedstock being used. Rapid cooling to 400–500°C is necessary to quench the reactions that destroy the more desirable products such as ethylene, propylene, and butadiene. The high-pressure steam generated in these exchangers is used to operate some of the gas compressors in the plant.

Coke formation is a problem in all TLXs, both in the cone of the TLXs and in the tubes. Coke increases the resistance to heat transfer so that less heat can be transferred (and less steam can be generated). Eventually, the thickness increases to a degree where the TLXs must be cleaned, usually manually. The reactor tubes and the TLXs must be designed and then operated such that the needs for decoking coincide for the two. As a rule, one TLX is provided for two reactor tubes.

The TLXs must be carefully designed to prevent excessive stresses in the shell or tubes when the TLX is heated to high temperatures. In addition, the design of the unit affects the amount of coke formed in the unit. Figure 6 illustrates a popular design, but others are also widely used.

**FIGURE 6** Transfer-line exchanger.

3. Product Gas Cooling and Compression

When the product gases exit from the TLX, they are still too hot for compression and must be cooled to essentially ambient temperatures. Part of the additional cooling is accomplished in some plants with a direct oil quench; the hot oil can be used to generate steam. Typically, in the final stage of cooling, a tower equipped with water sprays and a trap provides intimate contact between the water and the gaseous streams. The resultant water and liquid hydrocarbon phase (of heavier hydrocarbons) from this tower are separated in settling drums. Valuable products in both the water and oil phases are recovered and processed. The

water phase is often air cooled and recycled to the scrubbing tower. Part of the water phase is sent to a distillation tower for hydrocarbon recovery.

A typical pyrolysis plant employs large, complicated, centrifugal compressor systems driven by steam turbines, which require large volumes of high-pressure steam. Four- or five-stage centrifugal compressors are the norm with interstage cooling of the product stream. Liquid hydrocarbons and water formed in the cooling steps are separated from the gas phase after each stage. Care is taken to prevent high temperatures in the exhaust systems from any particular compression stage because of the propensity to form polymer deposits from butadiene or olefins. Temperatures of 110–120°C are upper limit guides.

4. Product Gas Treatment

The hydrocarbon gases that leave the compressor are usually subjected to three additional steps: (1) the removal of acid gas components, (2) the removal of acetylenic compounds, and (3) further water removal or drying.

The acid gases, usually CO₂ and H₂S, are removed by scrubbing with diethanolamine or monoethanolamine solutions and possibly an additional caustic treatment. Older processes used caustic solutions of 5–15 wt% NaOH followed by a water wash. Of course, the spent caustic creates a disposal problem; it must be neutralized with acid and then properly disposed of according to prevailing pollution and hazard waste standards. Different column configurations have been proposed, but usually large scrubber towers with well over 30 valve-type trays are used.

Although the acetylenic compounds, mostly acetylene itself, are present in relatively low concentrations (from 2000 ppm up to 33 wt%, depending on the feedstock and temperature of reaction), they must be removed to meet product specifications; most ethylene or propylene used in various petrochemical processes must be essentially free of acetylenic compounds (1–5 ppm). These compounds are removed when present in small amounts by hydrotreating in an adiabatic reactor utilizing catalysts of nickel-molybdenum or cobalt-molybdenum supported on a high-surface-area porous substrate such as alumina. A second unit of cleanup stage treatment using a much more sensitive and expensive palladium-supported catalyst follows. There are certain liabilities with this hydrogen treatment step. Some desirable compounds, including ethylene and butadiene, are partially hydrogenated. Care must be taken in the design of this hydrotreater and in its operation to minimize the loss of desirable products and to maintain the activity of the catalyst for extended periods. Operating variables of importance in this hydrotreater include temperature, partial pressure of hydrogen (which is always present in the product stream), and residence time. An al-

ternative to hydrogenation reactors is solvent scrubbing with dimethylformamide or acetone; the acetylene can be recovered in high yields.

The gaseous product stream is then dried to prevent ice formation during the refrigeration process. Usually, packed towers of alumina, silica gel, or molecular sieves are used. In spite of a higher initial cost, molecular sieves provide a number of advantages, including greater capacity for water and lower retention of heavy hydrocarbons. Multiple columns are utilized because a column will require regeneration approximately every 24 h. Regeneration is achieved by passing hot residue gases over the beds.

5. Product Separation

Low-temperature fractionation is the preferred method of hydrocarbon separation. Cascade refrigeration is used with propylene and ethylene, the commonly encountered refrigerants. Refrigeration compressors are large and have energy requirements essentially similar to those of the initial product gas compressors.

A demethanizer fractionation tower is frequently positioned first. This tower is often operated at $\sim 34 \times 10^5$ Pa, with temperatures low enough to obtain liquid methane. This tower is usually a tray-type column, although more recently packed towers have been introduced. The noncondensable gases (hydrogen, nitrogen, and carbon monoxide) and relatively pure methane can thus be separated from the C₂ and higher hydrocarbons.

The next unit in the separation stage is the deethanizer, which removes the C₂ hydrocarbons (ethylene and ethane) from the C₃ and higher hydrocarbons. The details of the separation train from the deethanizer through other units often vary significantly from plant to plant depending on which feedstock is utilized. For the general process described here, a depropanizer is the next unit in the train followed by a debutanizer. The former removes propane and propylene from the higher hydrocarbons and the latter removes essentially all of the C₄ hydrocarbons from the remaining C₅–C₈ hydrocarbons.

Separation of the C₂ stream to produce high-purity ethylene and ethane requires a large tower, sometimes the largest one in the plant. Separation of the C₃ stream to produce high-purity propylene and propane also requires a large tower, and in some plants it is the largest one. Separation of butadiene from the C₄ stream, if performed, is usually accomplished by extractive distillation. Aromatics are frequently recovered and separated to obtain benzene, toluene, and xylenes, especially when heavy feedstocks are used.

The distillation columns in these trains may be up to 13–18 m in diameter and 250 m high. Table VII presents typical variables for the distillation steps. The columns are

TABLE VII Information on Distillation Steps

Equipment	Overhead	Bottoms	Pressures ($\times 10^5$ Pa)	Reboiler (°C)	Reflux (°C)	Plates
Deethanizer	C ₂ hydrocarbons	C ₃ –C ₈ hydrocarbons	27	70–75	–10	40
Ethane-ethylene splitter	99.5–99.8% ethylene	Ethane	20	–5	–30	110
Depropanizer	C ₃ hydrocarbons	C ₄ –C ₈ hydrocarbons	16	105	50	40
Propane-propylene splitter	99.5% propylene	Propane	19	60	45	200

most often constructed with trays, although packed towers with demonstrated economies are now being introduced. The complex separation trains required include numerous heat exchangers to minimize energy demands. Selection of the best design and best operating conditions requires careful planning.

V. ECONOMICS

The designer of pyrolysis plants must consider numerous variables in order to achieve economic performance. The choice of a feedstock represents the most significant single variable. Plant location; plant size; and specific design details such as the separation train arrangement, compression and refrigeration configuration, and heat recovery contribute to the choices available to the designer. Table VIII lists typical contributions to the overall cost of operating an ethylene plant.

A. Feedstocks

There is often no single feedstock choice, since feedstock costs frequently vary erratically over a period of several years. A general guide to the influence of feedstock on capital investment of the entire pyrolysis for ethylene production is shown in Fig. 7. Net raw material costs for an ethylene plant often account for about 50–60% of the production costs, depending on whether the feedstock is a light material such as ethane or a heavier material such as naphtha.

B. Separation Train

The separations or fractionation equipment with its related exchangers may account for as much as 34% of the

capital investment. The total cost investment of this section is generally the single greatest figure. As discussed in the preceding sections, the orientation of the various separation towers, for instance, placement of the demethanizer before or after the deethanizer, choice of refrigerants, and economy of heat exchange, influences not only the capital investment, but also the operational costs. There has been a recent trend toward the use of packed columns rather than tray columns. The former often resulted in lower pressure drops, smaller towers, and/or lower reflux ratios.

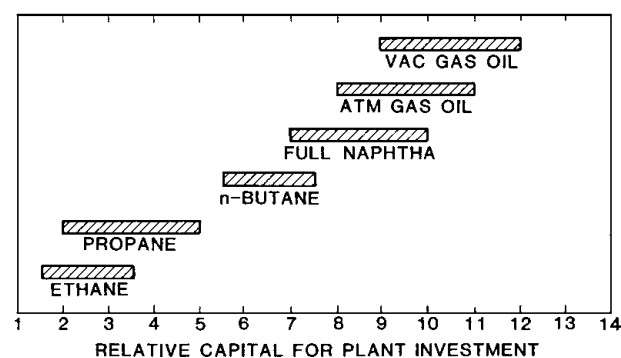
C. Cracking Reactor and Heat Recovery Equipment

About 30% of the total capital investment is the actual pyrolysis furnaces, provided that conventional reactors are used. Until nonconventional reactors fully demonstrate their economies, one cannot be certain how they will affect initial capital costs. Significant design considerations have been given to TLXs to recover the large quantities of energy available in the pyrolysis furnace exit gases other than the flue gases. Total utility costs, a significant part of which is for the fuel, frequently ranged from 22 to 27% of the production costs depending on whether the feedstock is ethane or naphtha.

In the United States, the preferred feedstocks for the production of ethylene and propylene continue to be lighter hydrocarbons such as ethane, propane, and their

TABLE VIII Production Costs

	Ethane feed (%)	Naphtha feed (%)
Raw materials	50–52	55–59
Utilities	24–26	21–23
Operational	5	4
Overhead	15–17	14–17

**FIGURE 7** Relative capital investments.

mixtures; they are still relatively plentiful and cheap. However, there is a continuing development of reactors that will accommodate the heavier feedstocks above naphtha. This includes gas oils and whole crudes. Although experiments have been conducted to assess pyrolysis kinetics and yield structures of heavy feedstocks such as coal liquids, shale oils, tar sand bitumen, and even waste plastics and biomaterials, commercialization of these complex mixtures will not take place until far in the future when the price of conventional feedstocks becomes prohibitive.

One highly evident development of feedstock supply is the movement of olefin production to sites of inexpensive feeds such as the Mid-East. There is a gradual shifting of the world's olefin production center of gravity toward these countries.

Since 1980 there has been a marked increase in temperatures, above 1000°C, with correspondingly shorter contact times in the high-temperature zone. Contact times in the millisecond region and lower are clearly targeted. These high-severity, low-contact-time reactors offer significant advantages in processing heavier feedstocks. Until there is a major breakthrough in new tube wall material for the conventional tubular furnace reactor, temperatures and contact times for the lighter hydrocarbon feedstocks will probably not change much.

SEE ALSO THE FOLLOWING ARTICLES

ACETYLENE • CATALYSIS, INDUSTRIAL • CHEMICAL KINETICS, EXPERIMENTATION • CHEMICAL THERMODY-

NAMICS • COMBUSTION • CRYOGENIC PROCESS ENGINEERING • PETROLEUM REFINING • RUBBER, SYNTHETIC • SYNTHETIC FUELS

BIBLIOGRAPHY

- Albright, L. F. (1985). "Processes for Major Addition-Type Plastics and Their Monomers," Chap. 2, Krieger, Melbourne, FL.
- Albright, L. F. (1988). *Oil Gas J.* August 15, 69–75; August 29, 44–48; September 19, 90–96; (1999). August 1, 35–40.
- Albright, L. F., and Baker, R. T. K. (1982). "Coke Formation on Metal Surfaces," ACS Symp. Ser. 202, Am. Chem. Soc., Washington, DC.
- Albright, L. F., Crynes, B. L., and Corcoran, W. H., eds. (1982). "Pyrolysis Theory and Industrial Practice," Academic Press, New York.
- Albright, L. F., and Marek, J. C. (1988). *Ind. Eng. Chem. Res.* **27**, 743–759.
- Baker, R. T. K., and Chludzinski, J. J. (1980). *J. Catal.* **64**, 464.
- Bergeron, M. P., Maharaj, E., and McCall, T. F. (March 1999). Spring National Meeting of the American Institute of Chemical Engineers, Houston, TX.
- (1999). *Chem. Eng. News*. July 15, 20–22.
- (2001). *Hydrocarbon Process.* March, 21–27.
- Kwilekar, A., and Bayer, G. T. (2001). *Hydrocarbon Process.* January, 80–84.
- Luan, T. C. (1993). "Reduction of Coke Deposition in Ethylene Furnaces," Ph.D. thesis, Purdue University, West Lafayette, IN.
- Schmidt, L. D. (1999). Personal communications, University of Minnesota.
- Szechy, G., Luan, T. C., and Albright, L. F. (1992). "Novel Production Methods for Ethylene, Light Hydrocarbons, and Aromatics," Chap. 18, pp. 341–360, Dekker, New York.
- Wysiekierski, A. G., Fisher, G., and Schillmoller, C. M. (1999). *Hydrocarbon Process.* January, 97–100.