McGRAW-HILL

ENCYCLOPEDIA OF SCIENCE & TECHNOLOGY

www.MHEST.com

19 U - ZYG

Ulcer

A lesion on the surface of the skin or a mucous membrane characterized by a superficial loss of tissue. Ulcers are most common on the skin of the lower extremities and in the gastrointestinal tract, although they may be encountered at almost any site. The diverse causes of ulcers range from circulatory disturbances or bacterial infections to complex, multifactorial disorders. The superficial tissue sloughs, leaving a crater that extends into the underlying soft tissue, which then becomes inflamed and is subject to further injury by the original offender or secondary infection.

Peptic ulcer. Peptic ulcer is the most common ulcer of the gastrointestinal tract and refers to breaks in the mucosa of the stomach or the proximal duodenum that are produced by the action of gastric secretions. Despite a plethora of clinical and experimental studies, it is still unknown why peptic ulcers develop. However, with rare exceptions, a person who does not secrete hydrochloric acid will not develop a peptic ulcer. Ulcers of the stomach tend to develop as a result of superficial inflammation of the stomach. These individuals tend to have normal or decreased amounts of hydrochloric acid. By contrast, most individuals with peptic ulcers of the proximal duodenum secrete excessive amounts of acid. Importantly, a bacterium, *Helicobacter pylori*, has been isolated from the stomach of most people with peptic ulcers and is thought to play a causative role.

The principal complications of peptic ulcers are hemorrhage and perforation, both of which may be fatal. Peptic ulcers do not ordinarily lead to cancer, but cancers of the stomach may become ulcerated and resemble peptic ulcers. Although stress has been anecdotally related to peptic ulcers for at least a century, serious doubt has been cast upon this concept. *See* HELICOBACTER; INFLAMMATION.

Ulcerative colitis. Ulcerative colitis is a disease of the large intestine characterized by chronic diarrhea and rectal bleeding. The disorder is common in the Western world, occurring principally in young adults. Its cause is not known, but there is some evidence for a familial predisposition to the disease. Psychosomatic explanations, which were popular at one time, are no longer thought to be credible. Most individuals have intermittent attacks, with periods of partial or complete remission. Those individuals with long-standing, extensive ulcerative colitis have a higher risk of colon cancer than the general population. Hemorrhage from the injured colon and perforation of the colonic wall are serious complications. Individuals with ulcerative colitis also tend to develop diseases of other organs, including the liver, joints, and skin. *See* CANCER (MEDICINE); HEMORRHAGE.

Other types. Other ulcers of the gastrointestinal tract are caused by infectious agents. Bacterial and viral infections produce ulcers of the oral cavity. Diseases such as typhoid, tuberculosis, and bacillary dysentery and parasitic infestation with ameba lead to ulcers of the small and large intestines.

Narrowing of the arteries to the legs caused by atherosclerosis, particularly in persons with diabetes mellitus, often causes ulcers of the lower extremities. Ulcers overlying severe varicose veins of the legs can also be troublesome. Such ulcers are difficult to heal unless the circulatory problems are brought under control. *See* ARTERIOSCLEROSIS; BACILLARY DYSENTERY; DIABETES; TUBERCULOSIS.

Emanuel Rubin

Bibliography. M. A. Peppercorn, *Contemporary Diagnosis and Management of Ulcerative Colitis and Proctitis*, 1995; S. Szabo and C. J. Pfeiffer, *Ulcer Disease: New Aspects of Pathogenesis and Pharmacology*, 1989.

Ultimobranchial bodies

Small, enigmatic structures which originate as terminal outpocketings from each side of the embryonic pharynx. They occur only in vertebrates, where they are almost universal but difficult to homologize.

They probably represent an expression of continued growth activity caudally, associated with pouch- or gill-forming potentialities of foregut entoderm. They are usually bilateral in mammals. The last (and sometimes next to last) "true" pharyngeal pouch and ultimobranchial primordium often unite to form a combined entity known as a caudal pharyngeal complex. During development in humans, the ultimobranchial bodies may be intimately related to the third, as well as the fourth, pharyngeal pouch. Here, as in most mammals, this complex would be the fourth or last.

Location and structure. Originally the bodies were regarded as vestigial lateral, or accessory, thyroids since, because of the mechanics of growth in mammals, they were found to join with lateral lobes of the thyroid. These "bodies" are now interpreted as relatively indifferent tissue, but capable of modification by various factors. When intimately incorporated within a growing and differentiating mammalian thyroid, they were frequently described as indistinguishably transformed (by embryonic induction) into thyroidlike tissue. They appeared to function especially in this way only during periods of thyroid activity such as would occur under the influence of the controlling hormone thyrotropin, but they also have added significance. *See* EMBRYONIC INDUCTION.

In lower vertebrates, ultimobranchial bodies do not join the thyroid but remain isolated in connective tissue between the thyroid gland and heart. Because of their position in fishes and amphibians, these structures were referred to as suprapericardial bodies.

Morphologically, ultimobranchial tissue can be highly variable, even within individuals of a single species. In lower forms, it is often vesicular, and evidence of secretory activity does occur, especially in elasmobranchs, reptiles, and birds. In mammals, it may reflect the secretory features more typical of lower animals. However, within old, atrophic or relatively inactive thyroids, the tissue is often cystic and nonsecretory. It can become mucus-secreting, and this may be reminiscent of its entodermal origin. Among mammals, depending upon the species, it has at various times been thought either to degenerate without trace, to transform by induction into functional thyroid follicles, or occasionally to persist after birth as variable, and multiple, epithelial cysts which sometimes stimulate a leukocytic reaction. Now established is the fact that, at least in sheep and rats, ultimobranchial tissue can occur almost universally as potentially dynamic, cystic metaplasias in postnatal thyroids, which may transform, under certain conditions, into active, centrifugally proliferating, stratified squamous epithelium, resulting in cellular aggregates (neoplasias).

Furthermore, because of the proximity of these ultimobranchial bodies to more cephalic "true" pouches (fifth, fourth, or third in ascending vertebrate phyla), which normally give rise to parathyroid and thymus tissue, and because foregut entoderm may possess reciprocal potentialities, this tissue on occasion may also carry induced, if not intrinsic, attributes for formation of accessory thymus or parathyroid tissue. *See* PARATHYROID GLAND; THYMUS GLAND.

Function and fate. Since in many animals the ultimobranchial bodies have been shown to be inconsistent and to possess few specific, permanent structural attributes, they have been regarded by many as not only variable and erratic in development but also relatively unstable after birth. Despite this unusual character for endocrine tissue, it has been discovered that the ultimobranchial bodies can produce a hormone distinctly different from that of other thyroid hormones. This hormone is a polypeptide and has been labeled calcitonin. Collectively, these cells constitute an endocrine system that plays a role as a delicate and perhaps subtle calcium-regulating hormone. In response to hypercalcemia (high calcium levels of blood), these ultimobranchial cells (or "C" cells) secrete large quantities of this hormone, which rapidly lowers plasma calcium by inhibiting bone resorption, thus preventing excessive osteolysis and calcium deposition in tissues. Although these bodies have been demonstrated in many lower vertebrates, ultimobranchial cells in the shark, chicken, dog, and sheep are especially rich in calcitonin. *See* HORMONE.

However, because of the morphologic instability of this "gland" within the mammalian thyroid (variable cystic changes), including that of humans, a functional instability, or lability, may be implied. It can be influenced by certain intrinsic and extrinsic factors. Not only is it subject to marked structural alterations experimentally, including susceptibility for tumorigenesis, by hormonal and dietary manipulations, such as vitamin A deficiency or carcinogens, but it is seemingly labile, normally, at least in sheep and rats. Possibly no other mammalian endocrine organ exhibits such marked structural variability and therefore, perhaps, subtle systemic dependence. It is vulnerable to change and may reflect changes in blood calcium.

John H. Van Dyke

Ultracentrifuge

A centrifuge of high or low speed which provides convection-free conditions and which is used for quantitative measurement of sedimentation velocity or sedimentation equilibrium or for the separation of solutes in liquid solutions.

The ultracentrifuge is used (1) to measure molecular weights of solutes and to provide data on molecular weight distributions in polydisperse systems; (2) to determine the frictional coefficients, and thereby the sizes and shapes, of solutes; and (3) to characterize and separate macromolecules on the basis of their buoyant densities in density gradients.

The ultracentrifuge is most widely used to study high polymers, particularly proteins, nucleic acids, viruses, and other macromolecules of biological origin. However, it is also used to study solution

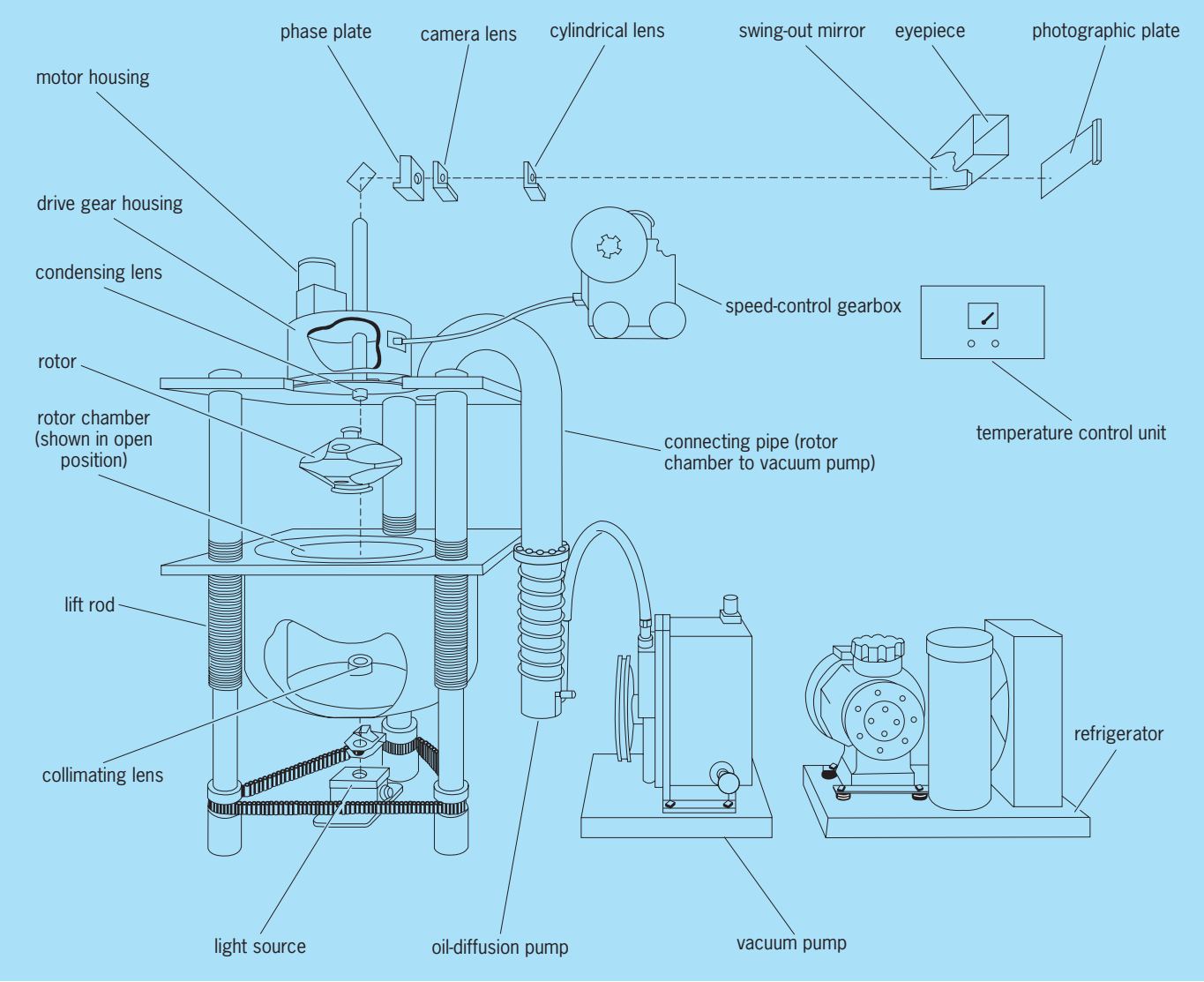


Fig. 1. Diagram of an electrically driven analytical ultracentrifuge. (*Beckman Instruments*)

properties of small solutes. In applications to macromolecules, one distinguishes between the analytical ultracentrifuge, which is used for accurate determination of sedimentation velocity or equilibrium, and the preparative ultracentrifuge, which is used to separate solutes on the basis of their sedimentation velocities or buoyant densities. *See* CENTRIFUGATION.

The application of a centrifugal field to a solution causes a net motion of the solute. If the solution is denser than the solvent, the motion will be away from the axis of rotation. The nonuniform concentration distribution produced in this way leads to an opposing diffusion flux tending to reestablish uniformity. In sedimentation-velocity experiments, sedimentation prevails over diffusion, and the solute sediments with finite velocity toward the bottom of the cell, although the concentration profile may be markedly influenced by diffusion. In sedimentation-equilibrium experiments, centrifugal and diffusive forces balance out, and an equilibrium concentration distribution results which may be analyzed by thermodynamic methods.

Instrument design. Although the first ultracentrifuges were oil-turbine- or air-driven, all modern instruments are electrically driven. The components of an electrically driven analytical ultracentrifuge are shown in **Fig. 1**. The rotor, cell, drive mechanism, and temperature control are designed to achieve high centrifugal accelerations and conditions of convectionless sedimentation. Special rotors have been designed for analytical and preparative purposes. An analytical rotor is shown in **Fig. 2**, and a preparative rotor in **Fig. 3**. Analytical cells, which contain the solution under study, are sector-shaped to allow unimpeded radial flow. The most commonly used analytical ultracentrifuge can be run at speeds between about 2000 and 60,000 revolutions/min. This fastest speed generates centrifugal forces of nearly 260,000 *g*. *See* CENTRIFUGAL FORCE.

Three types of optical systems are used in analytical ultracentrifuges to measure concentration

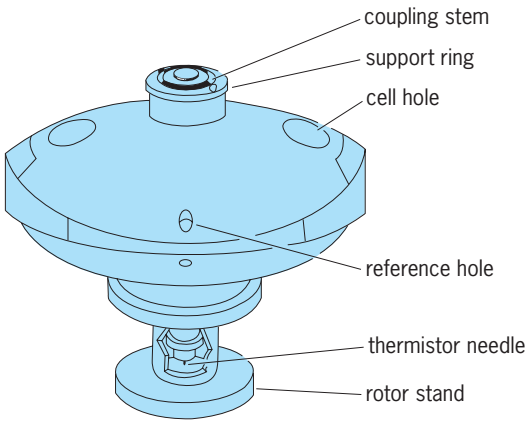


Fig. 2. Analytical rotor. (*Beckman Instruments*)

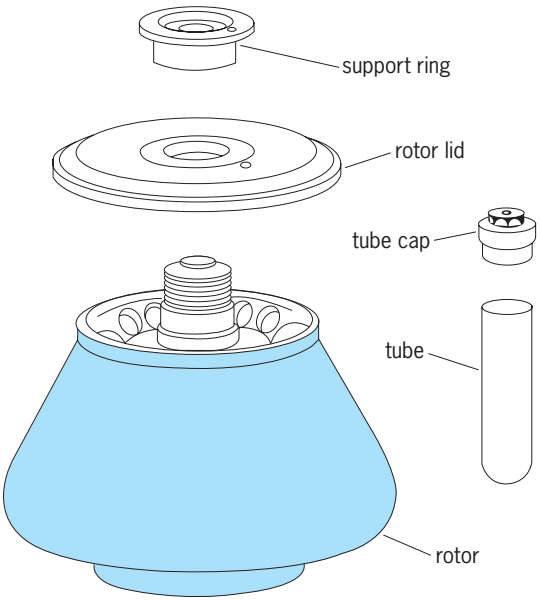


Fig. 3. Fixed-angle preparative rotor and tubes. (*Beckman Instruments*)

distributions while the instrument is in operation. The schlieren system measures refractive index gradients or, equivalently, concentration gradients dc/dr versus distance r from the axis of rotation. The Rayleigh system measures concentration c versus r by creation of interference fringes. The absorption optical system measures c versus r by visible or ultraviolet absorption by chromophoric solutes. Concentration distributions can also be determined in the preparative ultracentrifuge by optical absorption or by radioactivity of labeled solutes after sedimentation has ceased. *See* SCHLIEREN PHOTOGRAPHY.

Molecular weight determination. Four methods of molecular weight determination are available with the ultracentrifuge.

1. *Sedimentation equilibrium*. If equilibrium has been reached between sedimentation and diffusion, the relation between the solute concentration profile and molecular weight M is Eq. (1), where C_1 and C_2 are concentrations of solute at points r_1 and r_2 in the cell, $\overline{v}$ is the solute partial specific volume, ρ is the solution density, ω is the angular velocity, R is the gas constant, and T is the absolute temperature. For very large macromolecules, equilibrium can be attained only at low speeds, and then may require a day or more of continuous running to be established.

$$M = \frac{2RT \ln (C_2/C_1)}{(1 - \overline{v}\rho)\omega^2(r_2^2 - r_1^2)} \quad (1)$$

2. *Archibald approach-to-equilibrium method.* The condition for sedimentation equilibrium, that no net flux of solute occurs, is satisfied at the ends of the cell column throughout the approach to equilibrium. Therefore an equilibrium result for molecular weight may be obtained from the concentration distribution at either end of the solution column in much shorter times than required for complete equilibrium to be reached. The appropriate relation is Eq. (2).

$$M = \frac{RTrc}{(1 - \overline{v}\rho)\omega^2(dc/dr)} \quad (2)$$

3. *Sedimentation equilibrium in a density gradient.* Binary solutions of low-molecular-weight components of different densities, such as H_2O and CsCl, will exhibit a variation in density with distance from the axis of rotation at sedimentation equilibrium. When a macromolecular species is present in such a density gradient, it will take up a position in the cell where its buoyant density equals the density of the medium. The molecular weight of the macromolecule is related to the density gradient $d\rho/dr$, the standard deviation σ of the width of the band of macromolecules, and the position of the band center r, by Eq. (3).

$$M = \frac{RT}{\overline{v}\omega^2(d\rho/dr)\sigma^2 r} \quad (3)$$

4. *Sedimentation velocity.* This method utilizes centrifugal accelerations which are great enough to produce conveniently measurable sedimentation rates for solutes. A sedimentation coefficient S is obtained which is defined by Eq. (4), where dr/dt is

$$S = \frac{dr/dt}{\omega^2 r} \quad (4)$$

the sedimentation velocity of the solute. If the diffusion coefficient D of the solute is known, the molecular weight is obtained from relationship (5). The

$$M = \frac{RTS}{D(1 - \overline{v}\rho)} \quad (5)$$

measured sedimentation coefficient alone may be used to determine the molecular weight of a polymer if an S versus M curve has been calibrated with homologous polymers of known molecular weight. The calibration curve is summarized in the form of Eq. (6), where K and $a$are molecular-weight-

$$S = KM^a \quad (6)$$

independent parameters characteristic of a given polymer-solvent system.

Sedimentation velocity experiments are often run in density gradients, which stabilize the solution against convective disturbances.

If the solute is polydisperse, the molecular weights determined by the methods listed above will represent various averages over the molecular weight distribution. Most common of these are the number-average molecular weight M_n, which is the first moment of the distribution; the weight-average molecular weight M_w, which is the ratio of the second moment to the first; and the z-average molecular weight M_z, which is the ratio of the third moment to the second. By allowing the determination of these averages in a single experiment, sedimentation equilibrium is a valuable tool for characterizing molecular weight distributions in polymers. *See* MOLECULAR WEIGHT.

Frictional coefficient. Ultracentrifugation is used to characterize the sizes and shapes of solutes by determination of their translational frictional coefficient f, which is the proportionality constant between the force acting on a solute molecule and its velocity. The relation between sedimentation coefficient and frictional coefficient is given by Eq. (7), where N is the Avogadro number.

$$S = \frac{M(1 - \overline{v}\rho)}{Nf} \tag{7}$$

For a sphere of radius R, f is given by Stokes' law, Eq. (8), where η is the solvent viscosity. The frictional coefficient has also been calculated for prolate and oblate ellipsoids of revolution, rigid rods, flexible polymer chains, and chains of intermediate stiffness.

$$f = 6\pi\eta R \tag{8}$$

Buoyant density. Sedimentation equilibrium in a density gradient enables a solute to be characterized on the basis of its buoyant density, which is essentially the reciprocal of its solvated partial specific volume. This technique has been especially useful in the study of nucleic acids, where it has enabled distinctions between molecules to be made on the basis of base composition, deoxyribonucleic acid or ribonucleic acid, degree of isotopic substitution, double- or single-strandedness, and other important features. *See* DEOXYRIBONUCLEIC ACID (DNA).

Victor A. Bloomfield

Bibliography. W. Borchard, H. G. Killian, and G. Lagaly, *Progress in Analytical Ultracentrifugation*, 1992; F. Kremer and G. Lagaly (eds.), *Ultracentrifugation*, 1994; P. Sheeler, *Centrifugation in Biology and Medical Science*, 1981, reprint 1990; T. Svedberg and K. O. Pedersen, *The Ultracentrifuge*, 1940.

Ultrafast molecular processes

Various types of physical and molecular changes occurring on time scales of 10^{-14} to 10^{-9} s that are studied in the fields of photophysics, photochemistry, and photobiology. The time scale ranges from near the femtosecond regime (10^{-15} s) on the fast side, embodies the entire picosecond regime (10^{-12} s), and borders the nanosecond regime (10^{-9} s) on the slow side.

A typical experiment is initiated with an ultrashort pulse of energy—radiation (light) or particles (electrons). Rapid changes in the system under study are brought about by the absorption of these ultrashort energy pulses. Excited molecules, for example, may change their structure internally, may gain or lose protons or electrons, may react chemically, may exchange their energy with nearby molecules, or may simply lose memory of the initial dipole direction in which they were excited (dephasing). These changes can be measured by ultrafast detection methods. Such methods are usually based on some type of linear or nonlinear spectroscopic monitoring of the system, or on optical delay lines that employ the speed of light itself to create a yardstick of time (a distance of 3 mm equals a time of 10 ps). These studies are important because elementary motion, such as rotation, vibrational exchange, chemical bond breaking, and charge transfer, takes place on these ultrafast time scales. The relationship between such motions and overall physical, chemical, or biological changes can thus be observed directly. *See* CHEMICAL DYNAMICS; OPTICAL PULSES.

Ultrashort pulses. Just as a very fast shutter speed is necessary to obtain a sharp photograph of a fast-moving object, energy pulses for the study of ultrafast molecular motions must be extremely narrow in time.

Mode locking. The basic technique for producing ultrashort light pulses is mode-locking a laser. One way to do this is to put a nonlinear absorption medium in the laser cavity, which functions somewhat like a shutter. In the time domain, the developing light pulse, bouncing back and forth between the laser cavity reflectors, experiences an intensity-dependent loss each time it passes through the nonlinear absorption medium. High-intensity light penetrates the medium, and its intensity is allowed to build up in the laser cavity; weak-intensity light is blocked. This process shaves off the low-amplitude edges of a pulse, thus shortening it. One of the laser cavity reflectors has less than 100% reflectance at the laser wavelength, so part of the pulse energy is coupled out of the cavity each time the pulse reaches that reflector. In this way, a train of pulses is emitted from the mode-locked laser, with each pulse separated from the next by the round-trip time for a pulse traveling at the speed of light in the laser cavity, about 13 ns for a 2-m (6.6-ft) laser cavity length. *See* LASER; NONLINEAR OPTICS.

Dye lasers. The most convenient lasers for ultrafast molecular experiments are the mode-locked continuous-wave (CW) dye lasers. A jet stream of dye solution shaped like a thin plate is the laser gain medium. Because of the broad spectral characteristics of various dye solutions, these lasers can provide tunable wavelengths over an entire range, from the near-ultraviolet through the visible to the

near-infrared. Dye lasers are mainly of three types. In the first type, the gain jet is synchronously excited (pumped) by a mode-locked argon-ion laser or a frequency-doubled solid-state laser, such as neodymium/yttrium aluminum garnet (Nd:YAG). The pump pulse widths are usually around 70–100 ps, while the dye laser output pulses are 1–10 ps wide.

The second type of dye laser is a hybrid of the first type and a saturable absorber dye jet in the laser cavity, which reduces the laser pulse width. Subpicosecond pulses with discrete tunability can be generated with this type of dye laser.

The third type is the colliding-pulse mode-locking dye laser. This laser generates the shortest pulses, about 40 fs. The configuration is a ring cavity pumped by a continuous-wave ion laser. An absorber jet is placed at a quarter cavity length from the gain jet. Two counterpropagating pulses are adjusted to meet in the pulse-shaving saturable absorber. This introduces a fourfold greater light intensity in the absorber jet, which enhances the effectiveness of the saturable absorber.

Ultrashort pulse compression. To obtain still narrower widths, pulses can be sent through a length of single-mode optical fiber that induces a monotonic frequency sweep (chirp). The red spectral components are on the leading edge of the chirped pulse, and the blue components are on the trailing edge. The pulse is then rephased by an element such as a grating pair or a prism pair, bringing the leading and trailing edges back into coincidence. Laser pulses with widths of 6 fs, or only three or four optical cycles, have been obtained in this way. It is noteworthy that such pulses, because of the Heisenberg uncertainty principle, have a broad spectral distribution (about 3700 cm^{-1}), a property that is useful in ultrafast probe experiments. Also, ultrashort pulses with an arbitrary shape can be achieved by placing amplitude and phase filters between the rephased pair. *See* SPREAD SPECTRUM COMMUNICATION; UNCERTAINTY PRINCIPLE.

Amplification. In general, the ultrashort light pulses produced by continuous-wave mode-locking dye lasers are not intense enough for carrying out most time-resolved experiments. The peak power per pulse is only about 100 mW for 100-fs pulses. Femtosecond amplifiers have been developed that operate with a repetition rate near 100 kHz (high repetition rate) to 10 Hz (low repetition rate). One type of 10-kHz amplifier is a 1.2-mm-thick dye jet, pumped by a copper vapor laser. The pulse width of the copper vapor laser is about 15 ns and is electronically synchronized with the dye laser pulses. The dye laser pulses are set to multipass, usually six times, through the amplifier jet, the energy of an amplified single pulse reaching several microjoules (1 $\mu J \div 100$ fs = 10 MW). The amplifiers with a lower repetition rate usually consist of a three-stage flow dye cell, which is pumped by a frequency-doubled Nd:YAG regenerative amplifier. The pump pulses are about 100 ps wide with an energy of 50–100 millijoules per pulse. The energy of the final amplified dye laser pulse can reach more than 1 mJ, while preserving the 100-fs pulse width (peak power more than 10 GW).

Supercontinuum. By passing intense ultrashort dye laser pulses into a cell containing water, or almost any kind of transparent solution, a supercontinuum (white light) output is created that has the same pulse width as that of the dye laser. The supercontinuum is produced by a combination of nonlinear optical and molecular effects. The spectral distribution of the supercontinuum can be caused to span a wavelength range from 190 to 1600 nm. An ultrashort light pulse with such broad spectral coverage is very useful as the probe pulse for time-resolved pump/probe spectroscopy.

Ultrashort electron pulses. Electron pulses of short duration can be produced from microwave linear accelerators. Single micropulses of electrons are produced by electronically gating the electron gun, and they are further bunched by synchronously applied microwave fields. Pulses containing 10^{10}–10^{11} electrons with a time width as short as 5 ps can be produced in this way. The approximately 20-MeV energy per electron in the electron pulse is over a million times greater than the energy per photon in a laser pulse. However, the high energy is not necessarily an advantage when studying molecular events, which typically require no more than a few electronvolts for their initiation. The excess energy may cause a number of events to take place simultaneously, thus complicating interpretation. *See* PARTICLE ACCELERATOR.

Detection techniques. In addition to the extremely narrow energy pulses in the time domain, special ultrafast detection techniques are required in order to preserve the time resolution of an experiment.

Streak camera detection. A streak camera operates in some ways like an oscilloscope. A slit image of the light from the event is focused onto photocathode material deposited on a transparent substrate that is housed in an evacuated image tube. The resulting electron packet is accelerated and sent through a linearly falling electric field (voltage ramp) that can be set to decay on nanosecond-to-subpicosecond time scales. The resulting electron packet is thereby deflected onto a moderately persistent phosphor screen. The degree of deflection depends on the instantaneous value of the field—electrons in the packet that arrive early encounter a larger field and are deflected more than electrons arriving later. The resulting phosphor streak therefore contains the time-intensity information of the experiment.

Pump/probe detection. In pump/probe experiments, the time resolution is limited only by the pulse width of the laser. A pump pulse is absorbed, creating a nonequilibrium concentration of intermediates. The decay of intermediates can be detected by transient absorption spectroscopy using a weak probe pulse optically delayed in time after the arrival of the strong pump pulse.

The development of the pump/supercontinuum probe technique has drawn considerable attention. In this type of experiment, an ultrashort pulse

of white light is used as the probe. The major advantages of the supercontinuum are that time-dependent spectral changes can be studied over a broad range of wavelengths and that a more complete view of the transient absorption spectrum can be obtained for analysis.

Time-correlated single-photon counting. The concepts behind this type of detection are based on nuclear counting technology. The probability distribution for emission of a single photon after an excitation event yields the time-intensity distribution of all the emitted photons. In a typical experiment, an ultrashort laser pulse is split in two by a beam-splitting reflector. One pulse impinges on a fast photodiode to generate the start signal for a time-to-amplitude converter. The other pulse, sometimes frequency-doubled, excites the sample. The input into a microchannel plate photomultiplier is adjusted to detect less than a single fluorescence photon per laser pulse. The detected signal is amplified and constitutes a stop signal. The output of the time-to-amplitude converter thus provides a voltage that is proportional to the time delay between the start and stop pulses. A multichannel analyzer is used to build up a histogram of these time delays by using many pulses from a laser with a high repetition rate. The methods of time-correlated single-photon counting are limited by photomultiplier response and detector electronics, providing time resolution no better than about 20 ps. However, the high detection sensitivity is often an advantage. There are still many interesting experiments on the 20–1000-ps time scale both in chemistry and biology that can be carried out with this detection technique.

Correlation methods. Two light signals from the same ultrafast source (autocorrelation) or from two different sources (cross-correlation), when combined in a nonlinear medium such as a crystal of potassium dihydrogen phosphate, produce a two-photon effect. Since no electronic instrumentation is involved, correlation methods can provide time resolution into the femtosecond regime. A widely used autocorrelator design is the interferometer type. If one pulse is delayed with respect to another, the intensity of the two-photon signal will be small, but when the pulses enter the nonlinear medium more nearly simultaneously, the signal increases. An asymmetric correlator has also been developed by the addition of a linear dispersive medium in one of the interferometer arms. In this way, the autocorrelator can be transformed into a powerful diagnostic instrument capable of providing a complete description, in amplitude and shape, of the ultrashort laser pulse.

The simplest asymmetric correlator uses propagation through glass. The pulse is broadened in a known manner by the linear dispersion of the glass. The broadened pulse is then cross-correlated with the original ultrashort pulse. The intensity correlation yields information on the shape of the ultrashort pulse, while its phase can be extracted from interferometric cross-correlation. The reconstruction leads to a complete determination of the complex electric-field amplitude of the pulse broadened by the glass. Taking the Fourier transform of this pulse and multiplying it by the transfer function of the glass leads to the Fourier transform of the original pulse. *See* INTERFEROMETRY.

Light-gating upconversion. A related correlation method that has been widely used is known as fluorescence upconversion. This is the best technique for resolving fluorescence on ultrafast time scales. The laser pulse is split in two. One pulse is then frequency-doubled and used to excite a sample, causing it to generate the fluorescence. The second beam is optically delayed relative to the pump pulse, and is sent into a nonlinear crystal together with the fluorescence to produce a sum-mixing frequency signal. Only that portion of the fluorescence coincident with the gating pulse is detected. The cross-correlation signal of gating pulse and fluorescence is detected, recorded, and analyzed as a function of time delay between pump and gating pulses to obtain time-intensity information about the experiment.

Biological applications. An interesting experimental application of ultrafast spectroscopy is in the resolution of the early ultrafast processes in the complex chain of events responsible for animal vision. Though various species of animals have widely different outward characteristics, the physicochemical processes through which the various organisms respond to light are remarkably similar. The overall response of the eye to light may take as long as a few hundredths of a second, yet it has been found that events on femtosecond and picosecond time scales dominate the early response following the absorption of light in the retina. *See* VISION.

A closely related photobiological process, which has received considerable attention by spectroscopists studying the femtosecond regime, is light-driven transmembrane proton pumping in purple bacteria (*Halobacterium halobium*). Each link in the complex chain of molecular processes, such as an ultrafast photoionization event, is "fingerprinted" by a somewhat different absorption spectrum, making it possible to sort out these events by pump/probe methods of ultrafast laser spectroscopy. The necessity for use of fast primary events in nature concerns overriding unwanted competing chemical and energy loss processes. The faster the wanted process, the less is the likelihood that unwanted, energy-wasting processes will take place. *See* LASER PHOTOBIOLOGY; LASER SPECTROSCOPY; SPECTROSCOPY.

G. Wilse Robinson; Ningyi Luo

Bibliography. T. Kobayashi et al. (eds.), *Ultrafast Phenomena XIV: Proceedings of the 14th International Conference, Niigata, Japan, July 25-30, 2004*, 2005; M. M. Martin and J. T. Hynes (eds.), *Femtochemistry and Femtobiology: Ultrafast Events in Molecular Science*, 2004; R. A. Mathies et al., Direct observation of the femtosecond excited-state cis-trans isomerization in bacteriorhodopsin, *Science*, 240:777–779, 1988; A. Mozumder (ed.), Early events in radiation chemistry (special issue), *Rad. Phys. Chem.*, vol. 31, no. 1, 1989; I. Tanaka, I. N. Molin, and Mustafa El-Sayed (eds.), *Ultrafast Processes in*

Chemistry and Photobiology, Chemistry in the 21st Century Monograph, 1995; A. H. Zewail, *Femtochemistry: Ultrafast Dynamics of the Chemical Bond*, World Scientific Series in the 20th Century Chemistry, vol. 3, 1994.

Ultrafiltration

A filtration process in which particles of colloidal size are retained by a filter medium while solvent plus accompanying low-molecular-weight solutes are allowed to pass through. Ultrafilters are used (1) to separate colloid from suspending medium, (2) to separate particles of one size from particles of another size, and (3) to determine the distribution of particle sizes in colloidal systems by the use of filters of graded pore size.

Ultrafilter membranes have been prepared from various types of gel-forming substances. Unglazed porcelain has been impregnated with gels such as gelatin or silicic acid. Filter paper has been impregnated with glacial acetic acid collodions of varying strengths to produce filters of graded porosity.

Another type of ultrafilter membrane is made of thin plastic sheet with millions of tiny pores evenly distributed over it. Flow rates of liquids and gases through these membranes are very high because the pore volume is 80% of the total membrane volume and the pores proceed through the filter in a direct path. Nominal pore diameters range from 10 nanometers to 5 micrometers. *See* COLLOID; FILTRATION.

Quentin Van Winkle

Ultralight aircraft

A lightweight, single-seat aircraft with low flight speed and power, used for sport or recreation. Ultralights evolved from hang gliders, and the first models were crude, home-built modifications of the Icarus, a rigid-wing, biplane, tailless hang glider popular in the mid-1970s. Newer ultralights are designed strictly as motorized aircraft and cannot be flown as hang gliders. Ultralights are sold as kits, requiring from 50 to 300 h for construction. *See* GLIDER.

Design. There have been tremendous variety of ultralight airframe configurations and control systems (see **illus.**). Airframe types include powered weight-shift hang gliders, flying wings, canard designs, antique biplane replicas, and traditional, monoplane structures with conventional tail designs. Control systems can be either weight-shift systems, two-axis controls, or conventional three-axis controls. In weight-shift designs pilots must shift their weight, using a movable seat, to change the attitude. Two-axis designs use controls to the elevator and rudder only; there are no ailerons. Modern designs use three-axis control systems; weight-shift and two-axis control systems are no longer in production. Three-axis controls resemble those of a standard airplane and include either ailerons or spoilerons (spoiler-type systems used to make turns). Some more sophisticated ultralights are equipped with wing flaps, used to steepen approach profiles and to land at very slow airspeeds. *See* AIRFRAME; FLIGHT CONTROLS.

Powered parachutes are also widely used. In this design, the pilot is suspended beneath a square parachute canopy. The parachute's engine and propellers are situated directly behind the pilot. Counterrotating propellers are employed in order to cancel out any uncommanded turning due to torque effects. Steering is controlled by manipulation of parachute risers, as in a conventional sport parachute. *See* PARACHUTE.

Typically, ultralight airframes are made of aircraft-grade aluminum tubes covered with Dacron sailcloth. Areas of stress concentration are reinforced with double-sleeving or solid aluminum components. Most ultralights are cable-braced and use aircraft-grade stainless steel cables with reinforced terminals. Wingspans average 30 ft (9 m), and glide ratios range from 7:1 to 10:1, depending primarily on gross weight and wing aspect ratio. *See* AIRFOIL; ASPECT RATIO; WING.

Ultralight engines are lightweight, two-stroke power plants with full-power values in the 28–35-hp (21–26-kW) range. They operate on a mixture of gasoline and oil, and most transmit power to the propeller via a reduction or belt drive, a simple transmission that enables the propeller to rotate at a lower, more efficient rate than the engine shaft. *See* PROPELLER (AIRCRAFT); RECIPROCATING AIRCRAFT ENGINE.

Regulation. Prior to September 1982, ultralights were unregulated in the United States. At that time the Federal Aviation Administration (FAA) enacted a regulation that set performance and operational parameters for ultralights. The limitations include a maximum empty weight of 254 lb (115 kg), a maximum level flight speed of 63 mi/h (55 knots or 28 m/s), a stall speed no higher than 27 mi/h (24 knots or 12 m/s), a maximum 5-gallon (19-liter) fuel capacity, and single-seat occupancy. Operational limitations include provisions prohibiting flight at night and over congested areas and controlled airspace. Ultralight pilots must also observe all visual flight rules (VFR). Additionally, ultralights may be used only for sport or recreational purposes and cannot receive a foreign or United States airworthiness certificate.

Since the FAA does not officially recognize ultralights as aircraft but instead treats them as vehicles, it does not require that the pilot have a pilot certificate or the vehicle have an airworthiness certificate. Instead, the FAA has approved a voluntary program of the United States Ultralight Association (USUA) that registers ultralight pilots and vehicles. The USUA also endorses instructors to teach in two-seat, ultralight-type vehicles.

In the interest of safer training, the FAA has authorized the use of heavier, more powerful two-seat, ultralight-type vehicles for use in training prospective ultralight pilots. In order to meet training needs, these vehicles may have empty weights up to 496 lb (225 kg), stall speeds of 40 mi/h (35 knots or 18 m/s), maximum cruise speeds of 80 mi/h (70 knots or

Ultralight aircraft. (*a*) Pterodactyl Ascender, a canard-type aircraft with 35-hp (26-kW) engine (*Freedom Fliers, Inc.*). (*b*) FP-303, with all-wood, latticed construction (*Fisher Flying Products*). (c) "Hummer" aircraft, with three-axis controls and V-tail (*Maxair*). (*d*) Quicksilver GT, which uses three-axis controls and 28-hp (21-kW) engine (*Eipper Aircraft*).

36 m/s), and maximum fuel capacities of 10 gallons (38 liters).

If an aircraft does not conform to all criteria established for ultralights, it must be registered as an experimental-category aircraft with the FAA, and the pilot must possess at least a student pilot certificate.

The National Transportation Safety Board investigated a number of fatal ultralight accidents and noted a high rate of in-flight airframe failure and loss of control, compared to accident rates for FAA-certificated lightplanes. Government officials undertook an investigation of the level of regulation of ultralights. Consequently, ultralight flying activity took a dramatic downturn in the mid-1980s. By the 1990s, the combination of safer designs and training in two-seaters brought an increase in flying activity, and a reduction in the numbers of total accidents. The number of fatal accidents has remained at a rate of about 35 per year. *See* AIRPLANE.

Thomas A. Horne

Bibliography. A. Y. Berger and N. Burr, *Berger-Burr's Ultralight and Microlight Aircraft of the World*, 2d ed., 1986; C. Hughs, *The Ultralight Pilot's Flight Training Manual*, 1995; B. Millspaugh, *Ultralight Airman's Manual*, 1987.

Ultrasonics

The science of sound waves having frequencies above the audible range, that is, above about 20,000 Hz. Original workers in this field adopted the term supersonics. However, this name was also used in the study of airflow for velocities faster than the speed of sound. The present convention is to use the term ultrasonics as defined above. Since there is no marked distinction between the propagation and the uses of sound waves above and below 20,000 Hz, the division is artificial. In this article the emphasis is on instrumentation, engineering applications, analytical uses, and medical applications. *See* SOUND.

Ultrasonic Generators and Detectors

The earliest instruments for producing ultrasonic waves in air were the Galton whistle and the Hartmann generator. These devices produce sound waves by blowing a jet of high-pressure air from a narrow slit against a sharp metal edge. The Hartmann generator raises the velocity of the jet above that of the sound waves and in effect generates standing shock waves. *See* SHOCK WAVE.

Piezoelectricity and magnetostriction. Ultrasonic transducers have two functions: transmission and

reception. There may be separate transducers for each function or a single transducer for both functions. The usual types of generators and detectorsv for air, liquids, and solids are piezoelectric and magnetostrictive transducers. Quartz and lithium niobate ($LiNbO_3$) crystals are used to produce longitudinal and transverse waves; thin-film zinc oxide (ZnO) transducers can generate longitudinal waves at frequencies up to 96 GHz. Another class of materials used to generate ultrasonic signals is the piezoelectric ceramics. In contrast to the naturally occurring piezoelectric crystals, these ceramics have a polycrystalline structure. The most commonly produced piezoelectric ceramics are lead zirconate titanate (PZT), barium titanate ($BaTiO_3$), lead titanate ($PbTiO_3$), and lead metaniobate ($PbNb_2O_6$). Composite transducers are transducers in which the radiating or receiving element is a diced piezoelectric plate with filler between the elements. They are called "composite" to account for the two disparate elements, the piezoelectric diced into rods and the compliant adhesive filler. *See* MAGNETOSTRICTION; PIEZOELECTRICITY.

Polymer transducers. Polyvinylidene fluoride (PVDF) is the most popular piezoelectric polymer material for shock sensors and for microphones and other sound pickup devices. It acts efficiently from low frequencies up to the megahertz range. The transducer is formed as a thin film and is polarized by applying a high voltage while the film is mechanically stretched. *See* MICROPHONE.

Pulse systems. These have been used to measure properties of liquids and solids. A short burst of ultrasonic waves from the transducer is sent into the medium and is reflected. By timing the received pulse from the transducer with respect to the transmitted pulse, or by a phasing technique, accurate velocity measurements can be made. Such techniques have been used widely in measuring the elastic constants of small specimens. The attenuation also can be measured by the rate at which pulses decrease with distance transmitted, but careful consideration must be given to spreading loss and to the losses in the seals connecting the transducers to the specimens.

Solid dielectric transducer. This type of transducer is made by stretching a sheet of dielectric film (typically Myla®) that is metallized on one side across a metal backplate. The conducting surface and the metal backplate form a capacitor. The separation between the surface and the plate varies because ultrasonic waves that are incident upon the surface compress gases trapped between the film and the backplate. The change in separation results in a flow of charge. These devices are used as both projectors and receivers in the frequency range 40–1000 kHz. The same transducer is used as both the projector and the detector in a pulse-echo system that is frequently used to measure range for autofocusing cameras. *See* ELECTRET TRANSDUCER.

Optoacoustics. Modulated light absorbed by a fluid or solid results in local heating and generation of an acoustic signal. Pulsed lasers can generate a very short ultrasonic signal with significant energy in the frequency range of 1 MHz to 10 GHz. Such pulses are useful in studying properties of materials and locating defects. *See* PHOTOACOUSTIC SPECTROSCOPY.

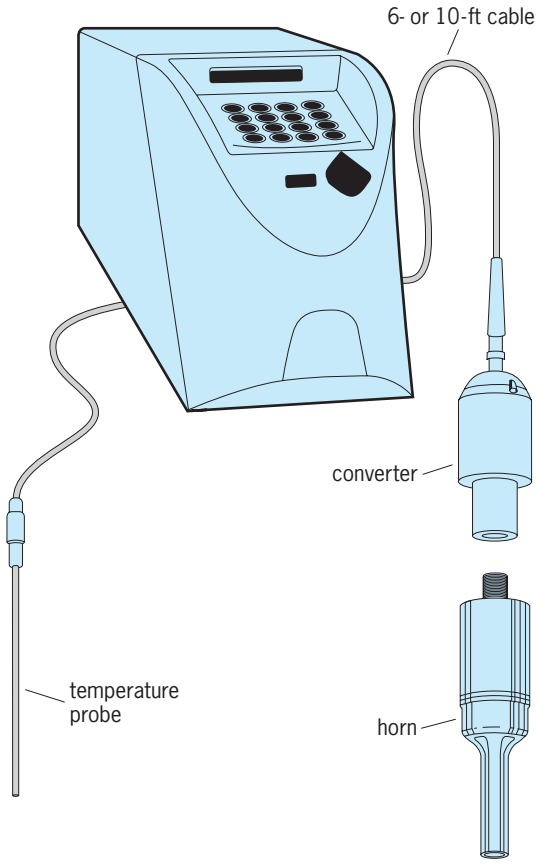


Fig. 1. Sonicator, used to produce high-intensity ultrasound. (*Misonix Inc.*)

High-power devices. High-power ultrasound (typically 600 W) can be obtained with sonicators, consisting of a converter, horn, and tip (**Fig. 1**). The converter transforms electrical energy to mechanical energy at a frequency of 20 kHz. Oscillation of piezoelectric transducers is transmitted and focused by a titanium horn that radiates energy into the liquid being treated. Horn and tip sizes are determined by the volume to be processed and the intensity desired. As the tip diameter increases, intensity or amplitude decreases.

Shear waves in liquids. A number of shear-wave transducers, most of them employing torsional or shear-wave generators of quartz, have been used to measure the shear viscosity and shear stiffness of liquids. These devices measure the acoustic resistance R and the acoustic reactance X of a plane shear wave sent into the liquid. *See* ACOUSTIC IMPEDANCE.

Measurements of this sort have shown that moderately viscous liquids have elastic as well as viscous properties, and they have been widely used in studying the motions possible in polymers and lubricating oils. A typical measurement of a polymer consisting of chlorinated biphenyls (aroclor series), shows that

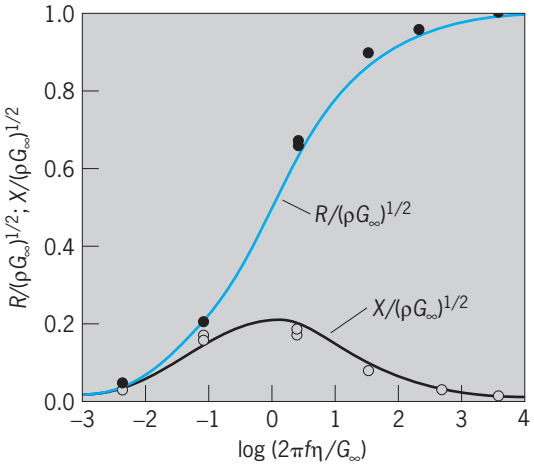


Fig. 2. Graph of $R/(\rho G_\infty)^{1/2}$ and $X/(\rho G_\infty)^{1/2}$ versus the logarithm of $2\pi f\eta/G_\infty$ for a polymer consisting of chlorinated biphenyls (aroclor series), where R is the acoustic resistance, X is the acoustic reactance, ρ is the density, G_∞ is the stiffness at infinite frequency, f is the frequency, and η is the viscosity of the liquid. Curves give theoretical values, and data points give measured values.

at low frequencies R and X are nearly equal, consistent with viscous-wave propagation (**Fig. 2**). For very high frequencies the reactance term tends to zero while the resistance term approaches a value given by Eq. (1), where ρ is the density and G_∞ the stiffness

$$R = (\rho G_\infty)^{1/2} \tag{1}$$

of the liquid at an infinite frequency. The frequency at which the reactance is a maximum (relaxation frequency) is given by Eq. (2), where η is the viscosity

$$f = \frac{G_\infty}{2\pi\eta} \tag{2}$$

of the liquid. For example, water with a stiffness of a porous solid of 10^8 newtons/m^2 (10^9 dynes/cm^2) and a viscosity of 10^{-3} pascal·s (0.01 poise) would have a relaxation frequency of 10^{10} Hz, which is above any present measurements. The viscoelastic properties of lubricating oils have been shown to contribute to the load-carrying capacity of spur gears operating at high speeds. *See* RHEOLOGY; VISCOSITY.

Engineering Applications

The engineering applications of ultrasonics can be divided into those dealing with low-amplitude sound waves and those dealing with high-amplitude (usually called macrosonics) waves.

Low-amplitude applications. Low-amplitude applications are in sonar (an underwater-detection apparatus), in the measurement of the elastic constants of gases, liquids, and solids by a determination of the velocity of propagation of sound waves, in the measurement of acoustic emission, and in a number of ultrasonic devices such as delay lines, mechanical filters, inspectoscopes, thickness gages, and surface-acoustic-wave devices. All these applications depend on the modifications that boundaries and imperfections in the materials cause in wave propagation properties. The attenuation and scattering of the sound in the media are important factors in determining the frequencies used and the sizes of the pieces that can be utilized or investigated. *See* SONAR.

Ultrasonic inspectoscopes. These transmit sound waves into a metal casting or other solid piece and determine the presence of flaws by reflections or by an interruption of the sound-wave transmission through the piece. Frequencies ranging from 500 kHz to 15 MHz are used. Such devices are among the best means for determining defects in metals, glasses, and ceramics, and they have also been applied in the inspection of automobile tires. *See* NONDESTRUCTIVE EVALUATION.

Ultrasonic thickness gages. These have been used in measuring the thickness of pieces when one side is not accessible, such as in boilers. Pulsing systems and continuous-frequency systems based on the resonance principle are used.

Surface-acoustic-wave devices. One application of ultrasonics involves the use of Rayleigh surface waves which were first studied in the case of earthquake waves on the surface of the Earth. In this mode the motion is along the surface with a small penetration into the interior. Such devices may be used in obtaining high-frequency wave filters (**Fig. 3**). Surface waves are generated by equally spaced interdigital electrode transducers. These generate a band of frequencies whose midfrequency is determined when the half Rayleigh wavelength equals the spacing of the transducer electrodes. Since the transducer sends out waves in both directions, the wave toward the back is absorbed by a resistive termination.

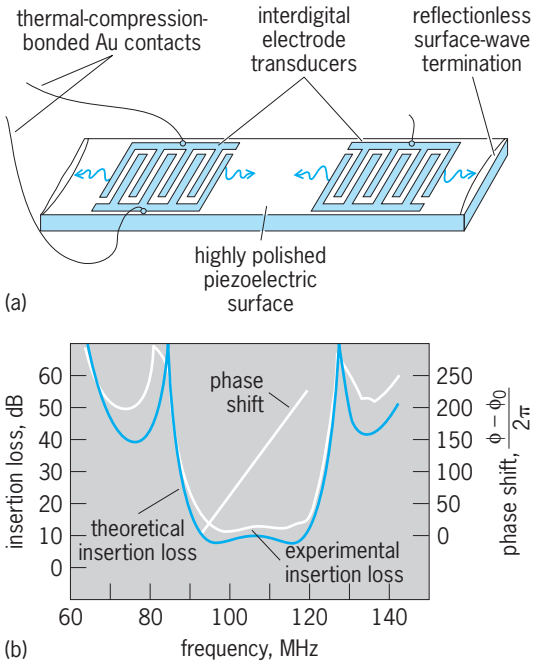


Fig. 3. Surface-wave transducers used as a bandpass filter. (*a*) Configuration. (*b*) Characteristics. Insertion loss is a form of attenuation. Phase shift is nearly linear. (*After E. A. Kraut, ed., Acoustic Surface Wave and Acoustooptic Devices, Optosonic Press, 1971*)

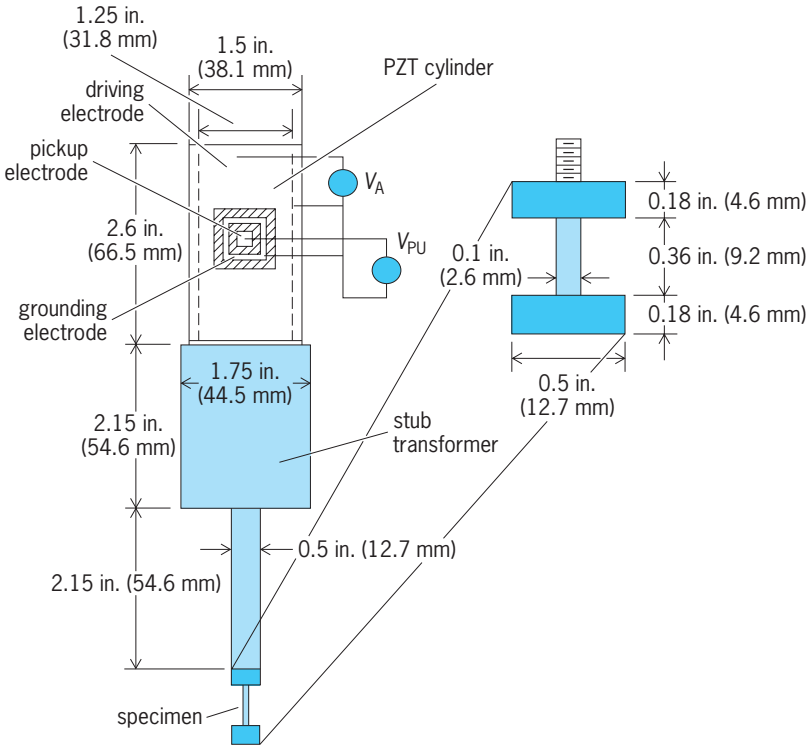


Fig. 4. Equipment for fatigue testing at ultrasonic frequencies, consisting of ultrasonic generator and specimen. V_A = applied voltage; V_{PU} = pickup voltage.

Most of the surface-wave devices use lithium niobate ($LiNbO_3$), which has a high electromechanical coupling and a low acoustic attenuation. For low-temperature coefficients, lithium tantalate ($LiTaO_3$) and S-cut quartz are sometimes used.

Surface-acoustic-wave devices are used for oscillators, nondestructive evaluation, and various processing devices. *See* SURFACE-ACOUSTIC-WAVE DEVICES.

Acoustic emission. An effect related to internal friction, motion of dislocations, and fatigue in materials is the noise in the specimens produced by strain. This is called acoustic emission. The first research on this phenomenon was done for rocks. This work started in the 1930s and was undertaken to test mine areas that were near danger regions for slides. Since the attenuation of sound waves in rocks is quite large at high frequencies, most of the measured frequencies were in the 150–10,000-Hz range. Later work has attempted to relate acoustic emission to earthquake properties. Acoustic emission and microfracturing in rocks may be directly related to the inelastic part of the stress-strain behavior, that is, to the internal friction. This suggests that the acoustic emission in rocks is connected with dislocation motion as has been established for metals.

High-amplitude applications. High-amplitude acoustic waves (macrosonic) have been used in a variety of applications involving gases, liquids, and solids. Some common applications are mentioned below.

Effects due to cavitation. A liquid subjected to high-amplitude acoustic waves can rupture, resulting in the formation of gas- and vapor-filled bubbles. When such a cavity collapses, extremely high pressures and temperatures are produced. The process, called cavitation, is the origin of a number of mechanical, chemical, and biological effects.

Cavitation plays an integral role in a wide range of processes such as ultrasonic cleaning and machining, catalysis of chemical reactions, disruption of cells, erosion of solids, degassing of liquids, emulsification of immiscible liquids, and dispersion of solids in liquids. Cavitation can also result in weak emission of light, called sonoluminescence, as discussed below. *See* CAVITATION; SONOCHEMISTRY.

Fatigue testing. Fatigue testing at ultrasonic frequencies has had considerable application because of the rapidity with which it can be accomplished. A typical device for testing (**Fig. 4**) consists of a PZT cylinder, vibrating at a frequency of the order of 20 kHz, attached to a mechanical transformer in the form of two quarter-wave stubs. The device is made from an alloy of titanium which has a low internal friction up to higher strains than any other materials. The stub transformer gives a transformation ratio proportional to the square of the ratio of the large diameter to the small diameter, in this case 12.25. The specimen, which consists of a thin shaft with two end weights, tuned to the same frequency as the transducer and the transformer, produces another factor of 5. Altogether, a linear strain of 6×10^{-3} can be produced if the material can withstand it.

By using such devices, it has been shown that the results obtained at high frequencies are closely similar to the results obtained at low frequencies, cycle for cycle. Constant total strain and constant displacement tests are possible. Positive mean load, elevated temperature, and corrosion ultrasonic fatigue tests can be performed routinely. The feasibility and advantages of performing crack propagation tests at ultrasonic frequencies have been demonstrated.

Other high-amplitude effects. Ultrasound is used widely in the cleaning of metal parts, such as in watches. The large acoustic forces actually break off particles and contaminants from metal surfaces. Ultrasound has been investigated for washing textiles.

One of the principal applications of ultrasonics to gases is particle agglomeration. This depends upon the fact that light particles can follow the rapid motion of the sound waves, whereas heavy ones cannot. Hence, light particles will strike and stick to heavy ones, reducing the number of small particles in the gas. The heavy particles eventually will fall to a collecting plate or can be drawn there by means of an electric field. This technique has been used in industry to collect fumes, dust, sulfuric acid mist, carbon black, and other substances.

Another industrial use of ultrasonics has been to produce alloys, such as lead-aluminum and lead-tin-zinc, that could not be produced by conventional metallurgical techniques. Shaking by ultrasonic means causes lead, tin, and zinc to mix.

Analytical Uses

In addition to their engineering applications, high-frequency sound waves have been used to determine the specific types of motions that can occur

in gaseous, liquid, and solid mediums. Both the velocity and attenuation of a sound wave are functions of the sound frequency. By studying the changes in these properties with changes of frequency, temperature, and pressure, indications of the motions taking place can be obtained.

Sound attenuation in fluids. In monatomic gases and monatomic liquids such as mercury, the sound attenuation can be explained as absorption due to viscosity and heat conduction. For such fluids, the attenuation A satisfies Eq. (3), where f is the frequency,

$$A = \frac{2\pi^2 f^2}{\rho v^3}\left[\frac{4}{3}\eta + \frac{(\nu - 1)K}{C_p}\right] \tag{3}$$

ρ the density, v the sound velocity, η the coefficent of viscosity, ν the ratio of specific heats, K the thermal conductivity, C_p the specific heat at constant pressure, and A the attenuation in nepers per meter, if the other quantities are in SI units.

Polyatomic liquids show additional attenuation due to relaxations of two types. Thermal relaxations, which have been demonstrated for gases and nonassociated liquids, that is, liquids that contain nonpolar molecules, occur by an interchange of energy between the longitudinal sound wave and the rotational and internal modes of motion of the gas or liquid molecules.

Structural relaxations occur for associated liquids, for polymer liquids, and also for solids. These relaxations take place when one part of the molecule moves from one position to another under the combined effect of the thermal- and sound-wave energy. A definite structure, such as that which occurs in associated liquids and polymer liquids, is required. *See* SOUND ABSORPTION.

Effects in solids. For solids, a variety of effects cause attenuation and velocity dispersion. Probably the simplest of these are thermal effects.

Thermal effects. When a solid body is compressed by an acoustic wave, the compressed part becomes hotter and the expanded part cooler. Thermal energy is transmitted from the hot part to the cool part. Since this energy comes from the acoustic wave, a loss or attenuation of the wave results which is proportional to the square of its frequency. For bars in flexural vibration, the thermal path is quite short, and the effect produced is large. Below a frequency f_0, determined by Eq. (4*a*), such a source produces an internal friction $1/Q$ given by Eq. (4*b*). Here K is the thermal con-

$$f_0 = \frac{\pi K}{2C_p \rho W^2} \tag{4a}$$

$$\frac{1}{Q} = \frac{Y_0^\sigma - Y_0^\theta}{Y_0^\theta}\left(\frac{ff_0}{f^2 + f_0^2}\right) \tag{4b}$$

ductivity, C_p the specific heat at constant pressure, f the frequency of the sound wave, ρ the density, W the width of the bar in centimeters, and Y_0^σ and Y_0^θ the adiabatic and isothermal values of Young's modulus, respectively. The velocity increases as a function of frequency (**Fig. 5**), while a corresponding internal friction occurs. If velocity and sound attenuation are graphed against frequency (Fig. 5), the shape of the resulting curves is the same in any medium with a single relaxation, although the horizontal and vertical scales vary from one medium to another.

The thermal path l for a longitudinal wave becomes smaller as the frequency of vibration increases. It is given by Eq. (5) where v is the velocity

$$l = \frac{v}{2f} \tag{5}$$

of propagation. Inserting this expression in Eq. (4*a*) with $l = W$, the relaxation frequency for a longitudinal wave is then given by Eq. (6). Above this frequency, the material is isothermal, whereas below f_0 it is adiabatic. This frequency is above 10^{10} Hz for most materials. The attenuation for a longitudinal wave for this thermoelastic effect is given by Eq. (7), where γ is the Grüneisen constant, B the bulk modulus, α the thermal expansion coefficient, C the specific heat per unit volume, ω is $2\pi f$, ρ the density of the medium, V_l the longitudinal velocity, T the absolute temperature in kelvins, and A the attenuation in nepers per meter, if the other quantities are in SI units. This source of attenuation is quite large for metals but provides only about 4% of the thermal attenuation for insulators.

$$f_0 = \frac{C_p \rho v^2}{2\pi K} \tag{6}$$

$$A = \frac{\omega^2}{2\rho V_l^3}\left(\frac{\gamma^2 KT}{V_l^2}\right) \qquad \gamma = \frac{3B\alpha}{C} \tag{7}$$

The main thermal attenuation is provided by the Akheiser effect. This loss is determined by the thermal conductivity and the nonlinear third-order elastic moduli.

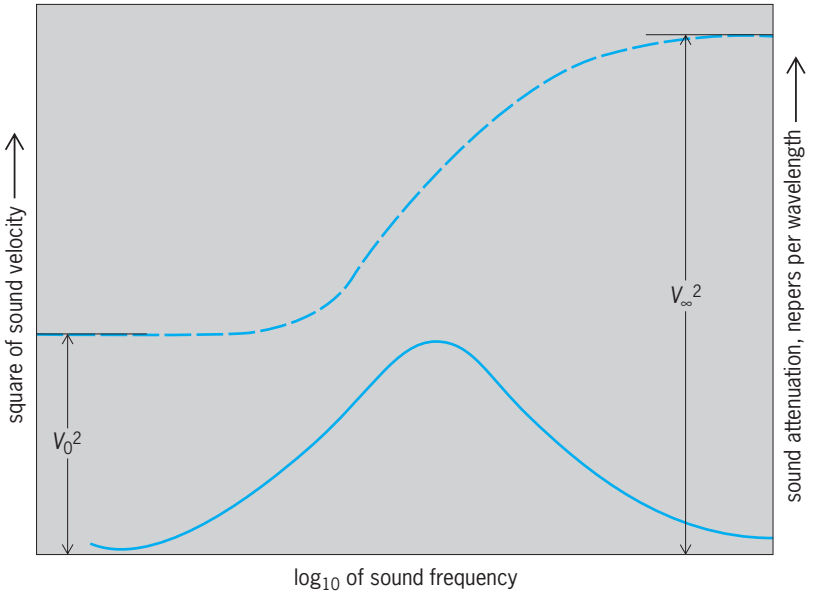


Fig. 5. Velocity dispersion (broken line) and corresponding attenuation per wavelength peak (solid line) for a medium with a single relaxation. The velocity increases as a function of frequency. One neper equals 8.7 dB.

Other relaxations. A number of relaxation phenomena are associated with the motion of impurity atoms, grain boundaries, domain boundaries, and other motions occurring in a solid. Interstitial atoms, such as nitrogen and carbon in iron, can cause an appreciable acoustic loss. These impurity atoms have preferred positions between the iron atoms in the crystal lattice. When a sound wave stretches the lattice in one direction and compresses it in a direction perpendicular to the first, the interstitial atoms, actuated by thermal energy, tend to go to the most open regions. When a compression due to the sound wave occurs, the reverse motion takes place. Since it requires a thermal activation energy H to move the impeding atoms aside, the frequency of jumping f follows Eq. (8), where f_0 is the frequency of vibration

$$f = f_0 e^{-H/(RT)} \tag{8}$$

of a nitrogen atom due to thermal motion ($\cong 10^{13}$ Hz), R the energy necessary to increase the temperature of 1 mole of atoms (6.022×10^{23} atoms) by 1°C, and T the temperature in kelvins. Since H is about 16,400 cal/mole (68,600 joules/mole) for nitrogen and R is 2 cal (8.31 joules), the relaxation frequency for this process is about 1 Hz at room temperature.

Other relaxations involving substitutional atoms have been observed at higher temperatures, since the substitutional atoms in this case have higher activation energies. Relaxations involving the rotation of grains in polycrystalline samples have been observed at high temperatures and low frequencies.

Much faster relaxations occur in magnetic processes involving the motion of domain walls in magnetic materials. A demagnetized specimen is made up of a number of domains within which the direction of magnetism is the same. Domains with directions of magnetism at right angles to or at 180° from the original direction are separated by regions called Bloch walls, in which the direction of magnetism changes from one domain to the other by small steps in the orientation of magnetism. A compressive stress in the same direction as the magnetic flux—for a positive magnetostrictive material—causes the domain to shrink, whereas it causes domains directed at 90° to expand. Hence, the domain wall moves as the stress changes from compressive to extensional. For a discussion of Bloch walls. *See* FERROMAGNETISM.

The domain walls can be held up by dislocations and other imperfections in the magnetic material, and a definite magnetic field or stress is required before the domain wall moves at all. On the reverse cycle the domain wall lags behind the applied stress (magnetic or elastic). The effect produces a hysteresis loop in the material and an acoustic loss called the microhysteresis effect. As the direction of magnetism changes, eddy currents are generated. These limit the velocity with which a domain wall can move and produce an acoustic loss for alternating stresses called the microeddy current effect. For a given size domain, there is some frequency for which the velocity is only half as large as that for low frequencies for the same applied magnetic field. The loss at this frequency is a maximum, and hence this frequency is a relaxation frequency. It can be shown that this frequency is determined by Eq. (9), where R is the

$$f_0 = \frac{R}{96\chi_0 l^2} \tag{9}$$

electrical resistivity of the material, χ_0 the initial magnetic susceptibility for a demagnetized material, and l the thickness of a domain. For nickel, for example, this frequency is of the order of 10^5 Hz. *See* DOMAIN (ELECTRICITY AND MAGNETISM); EDDY CURRENT; HYSTERESIS.

Many other relaxations occur, depending on the nature of the solid-state motion that can take place in the material. Ultrasonic measurements carried out over wide frequency and temperature ranges are powerful tools for investigating such motions.

Low-temperature data. Ultrasonic waves have provided significant information on processes that occur at temperatures near absolute zero. In liquids the most important results have been obtained for liquid helium, while for solids results have been obtained with metals at low temperatures which reveal a considerable amount of information about the mechanism of superconductivity.

Liquid helium. When helium is liquefied at its boiling point (4.2 K or −452°F) and cooled further, at the so-called lambda (λ) point (2.2 K or −455.7°F) there is a transformation of normal helium into superfluid helium. At the λ point there is an ambiguity in the sound velocity and also a high attenuation. Superfluid helium has a zero viscosity and a high thermal conductivity. The former leads to a small acoustic attenuation for normal sound, while the latter leads to the capability of transmitting thermal waves, the so-called second sound. Second sound can be initiated and detected by thermal means, and it has been found that the velocity is zero at 2.2 K (−455.7°F), rises to a maximum of 65 ft/s (20 m/s) at 1.7 K (−456.6°F), and decreases thereafter at lower temperatures. The velocity of normal sound varies from 750 ft/s (230 m/s) near absolute zero to 600 ft/s (180 m/s) near 2.2 K (−455.7°F). *See* LIQUID HELIUM; SECOND SOUND.

Attenuation at low temperatures. At very low temperatures the ultrasonic attenuation of pure normally conducting metals becomes high. For example, above 10 K (−442°F) the ultrasonic attenuation of pure tin (**Fig. 6**) is still relatively small and increases as the square of the frequency; whereas at 4 K (−452°F), at which temperature tin is still in the normal state, the attenuation is high and increases in proportion to the frequency. It has been shown that the added attenuation in the normal state is due to the transfer of momentum and energy from the acoustic wave to the free electrons in the metal. If the acoustic wavelength is greater than the electronic mean free path, this transfer determines an effective viscosity, and the attenuation increases in proportion to the square of the frequency. When the mean free path becomes longer than the acoustic wavelength, as it does at low temperatures, the energy communicated to the

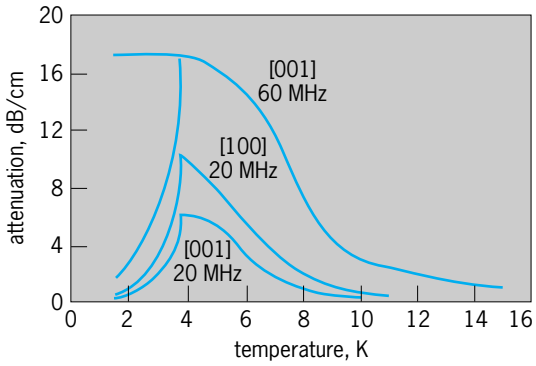


Fig. 6. Longitudinal sound-wave attenuation measurements for a single crystal of tin along the (001) axis and along the (100) axis. °F = (K × 1.8) − 459.67; 1 dB/cm = 2.5 dB/in.

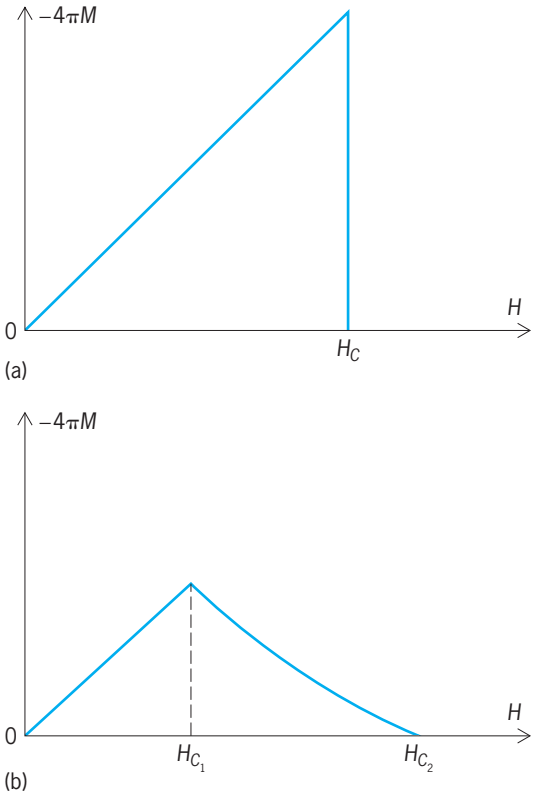


Fig. 7. Magnetization curves of long cylinders of type I and type II superconductors. The applied field *H* is directed along the axis of the cylinder.

electrons is not returned to the acoustic wave and a high attenuation results. The attenuation is proportional to the number of times the crystal lattice vibrates and hence to the frequency. *See* FREE-ELECTRON THEORY OF METALS.

As the temperature drops below the temperature at which tin becomes superconductive (3.71 K or −453.0°F), this source of attenuation drops rapidly to zero. The form of the temperature dependence (Fig. 6) has been used to confirm the Bardeen-Cooper-Schrieffer energy-gap theory of superconductivity. However, at lower frequencies, that is, from 10 to 100 MHz, losses due to dislocations can occur. These are different for the normal and superconducting states, and this difference has to be taken account of in order to determine the form of the energy-gap relation. For frequencies above 100 MHz, the attenuation due to dislocations is small compared to the electron-phonon loss, and direct measurements give the shape of the energy-gap curves. *See* SUPERCONDUCTIVITY.

Acoustic measurements are also useful for type II or high-field superconductors (HFS). For these types of superconductors, which are used for superconducting magnets, there are two critical fields, rather than the single field of type I superconductors (**Fig. 7**). In type I the magnetic flux is completely excluded from the interior of the material below H_c. For type II superconductors the magnetic flux is completely excluded from the interior only below H_{c1}. Between H_{c1} and H_{c2} the magnetic flux consists of flux vortices in the form of filaments directed along H, embedded in a superconducting material. When a dc electric current flows in a direction normal to H, each vortex experiences a force normal to its length, which causes it to move. The vortices are pinned by defects, and a finite current density is required before the vortices move. An alternating current or alternating stress causes motions of the pinned vortices which lag behind the applied forces. The result is an acoustic attenuation (**Fig. 8**). A sharp dip in the attenuation occurs near the superconducting field H_{c2}. Above H_{c2} the material is in the normal state and the attenuation rises rapidly with the field. *See* LOW-TEMPERATURE ACOUSTICS.

Resonant ultrasound spectroscopy. Resonant ultrasound spectroscopy (RUS) is an experimental technique for obtaining a complete set of elastic constants. It is based on the measurement of the free mechanical resonances of a sample of well-defined

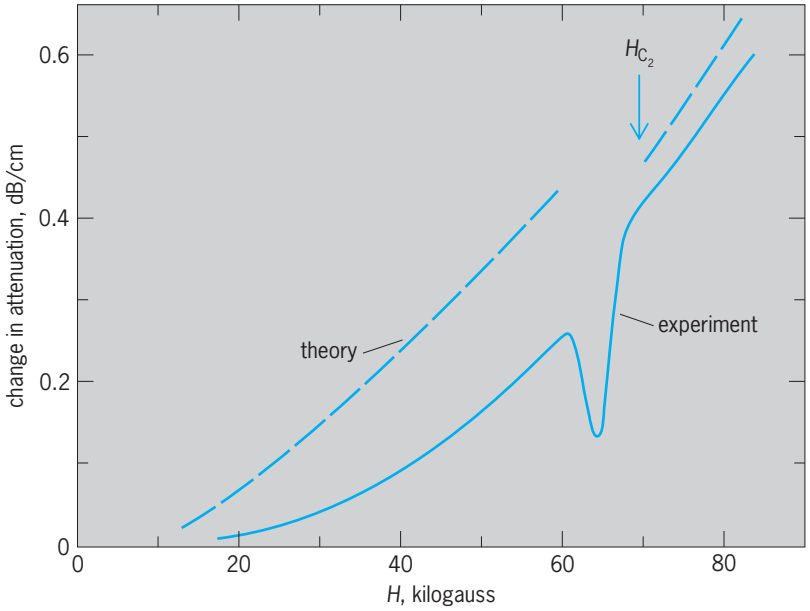


Fig. 8. Magnetic field variation of the attenuation of 9.1-MHz shear waves parallel to *H* in annealed Nb–25% Zr with temperature at 4.2 K (−452°F). 1 kilogauss = 0.1 tesla; 1 dB/cm = 2.5 dB/in.

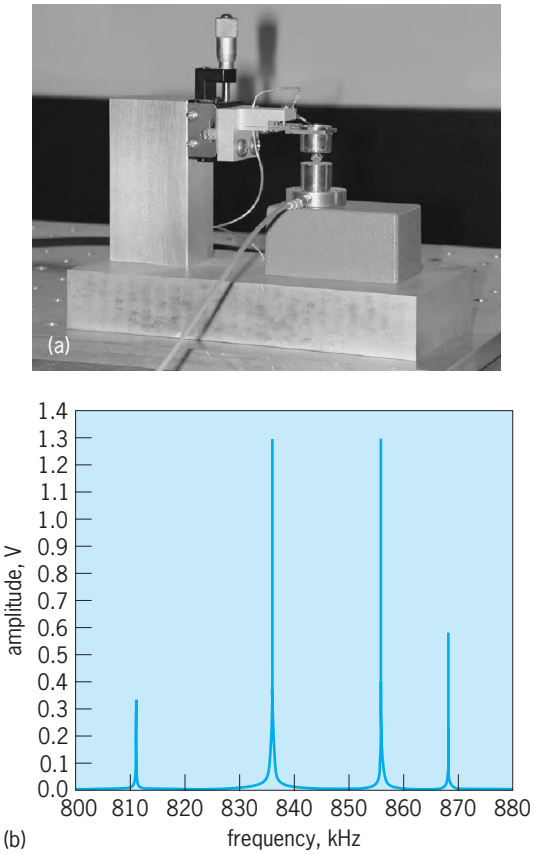


Fig. 9. Resonant ultrasound spectroscopy (RUS). (*a*) Transducers with mounted sample. (*Dynamic Resonance Systems Inc.*) (*b*) Part of a typical response of a scan.

shape, usually a sphere or rectangular parallelepiped. Given a sample's density and elastic constants, a spectrum of resonant frequencies can be calculated for free boundary conditions. It is the job of RUS to measure these resonant frequencies and fit the calculated resonances to the experimentally measured ones by adjusting the starting values for the elastic tensor.

In an RUS experiment, a specimen, typically 1–3 mm (0.04–0.12 in.) on a side, is held between two transducers with a minimal amount of pressure (**Fig. 9***a*). One transducer is used to excite the sample while the other measures its response. A few of the sharp resonances that are typically observed are shown in Fig. 9*b*. RUS does not require bonding of different transducers on the sample, as do conventional techniques, and all moduli are determined simultaneously. RUS is also sensitive to the symmetry of the object, so that certain symmetry-breaking effects (such as asphericity) are easily detected.

Sonoluminescence. Sonoluminescence is a phenomenon in which pulses of light are emitted by collapsing cavitation bubbles. There are two classes of sonoluminescence: multiple-bubble and single-bubble. Multiple-bubble sonoluminescence is associated with cavitation processes that involve many bubbles in relatively gassy liquids. Multiple-bubble cavitation is random in nature. As a result, the location and timing of multiple-bubble sonoluminescence events are largely unpredictable. Such conditions often arise in sonochemistry and industrial ultrasonic applications.

In single-bubble sonoluminescence, sonoluminescent pulses are emitted from a single, isolated bubble, trapped at the pressure antinode of a high-amplitude acoustic standing wave set up in a fluid-filled acoustic resonator. The host fluid is degassed. The bubble is driven into large-amplitude, nonlinear, radial oscillations by the levitation field. One sonoluminescent pulse is emitted in each acoustic cycle. Typical frequencies are in the range of 20–60 kHz. Single-bubble sonoluminescence can be generated with very simple apparatus and is easily visible to the unaided eye. The duration of the sonoluminescent pulse is estimated to be less than 50 picoseconds. There has been no direct measurement of the pulse duration. The emission can be extremely stable and repetitive, lasting for hours. It can also be quasiperiodic and chaotic, depending on the driving conditions. The spectrum is broad-band and increases in intensity in the ultraviolet. Slight cooling of the host liquid significantly increases the intensity. The intensity of the luminescence is sensitive to changes in the composition of gases within the bubble of less than 1%. Individual single-bubble sonoluminescence pulses are isotropic and unpolarized. Single-bubble sonoluminescence has been observed only in water and water solutions.

There is no completely satisfactory explanation for this phenomenon. The most plausible explanation is that during the highly spherically symmetric collapse of the isolated bubble, a shock wave is launched toward the interior of the bubble. When the shock wave reaches the center of the bubble, the temperature and pressure of the gas behind the shock front are both very high, leading to conditions under which light is generated. The short duration of the emission is accounted for by the fact that the conditions necessary for emission are quickly quenched as the shock wave diverges after focusing at the bubble's center.

Medical Applications

Application of ultrasonics in medicine can be generally classified as diagnostic and therapeutic. The more common of these at present is the diagnostic use of ultrasound, specifically ultrasonic imaging.

Ultrasonic imaging. The use of medical ultrasound procedures and equipment surpasses that of any other medical imaging modality. Three common imaging modes have evolved using ultrasound technology in medical imaging.

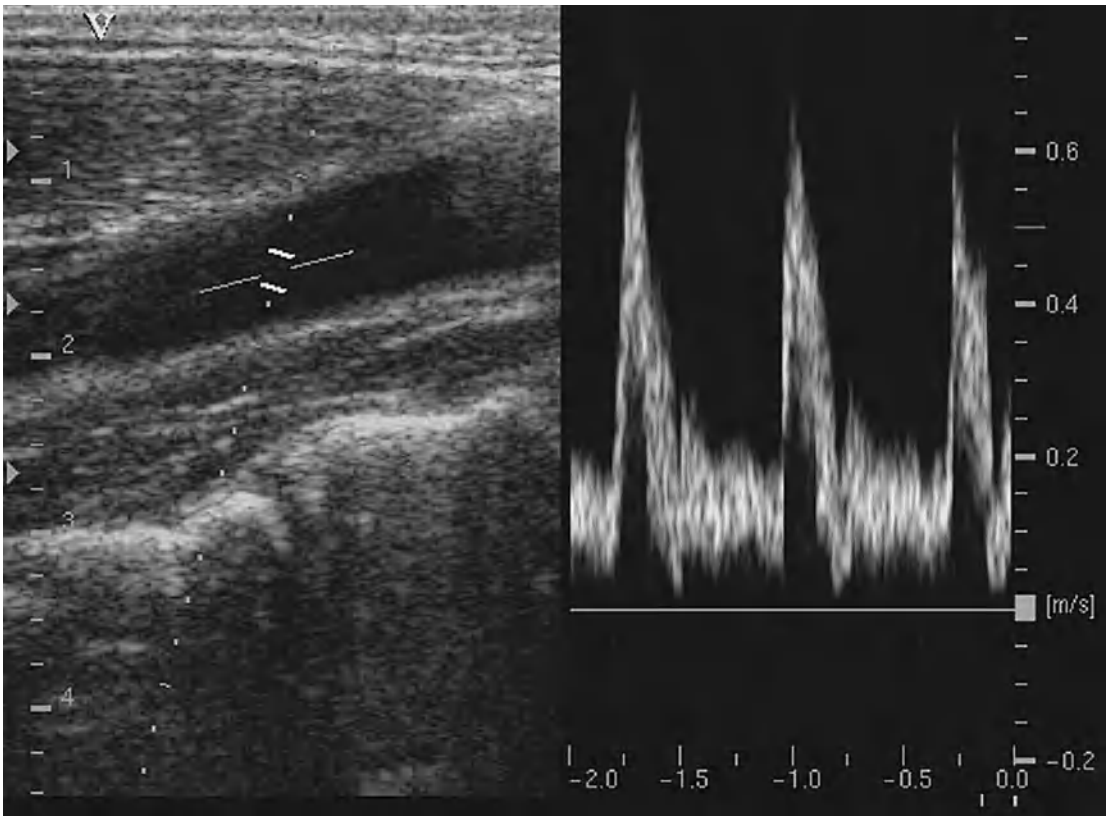
B-mode. Ultrasonic pulses are partially reflected from interfaces where a change occurs in the specific acoustic impedance. For pulse-echo imaging, an ultrasonic transducer, operating by the piezoelectric effect, emits an acoustic pulse and converts the returning echoes to voltage pulses whose amplitudes are proportional to the echo strength. The amplitude of voltage pulses can be used to modulate the intensity of the oscilloscope trace to create a B-mode

scan line. A two-dimensional B-mode image can be produced by combining a number of B-mode scan lines from different directions or positions into an image. The spatial relationship between the scan lines is preserved by recording the direction of the ultrasound beam and the range to targets, based on an assumed constant sound speed. B-mode images with more than 100 scan lines are the basis for many imaging systems used today, and video displays of digital data have replaced the simple oscilloscope traces used in early devices. The display can be updated at over 100 images per second.

Arrays of elements are used to focus and steer ultrasonic beams rather than fixed lenses and mechanical motion of an ultrasound transducer. Array systems include both linear arrays, where groups of elements are combined in a sequential fashion to scan an ultrasonic beam in a plane, and phased arrays where phasing of elements is used to steer the beam in the imaging plane. Phasing of elements is also used to focus the beam at multiple locations during repeated transmits or to continuously focus receive signals by phase coherent summation. Electronic focusing has been predominantly one-dimensional, allowing focusing within the imaging plane and orthogonal to the propagation direction. However, some systems have electronic control of the out-of-plane beam characteristics.

M-mode. Motion-mode or M-mode images are produced by combining B-mode scan lines based on their temporal relationship instead of the spatial relationship used in B-mode images. A sequence of B-mode scan lines is acquired with the same beam position and orientation, and is displayed vertically on a video screen with the most recent scan line on the right-hand side. As the next acoustic pulse is recorded, previous measurements are shifted to the left on the screen. The effect is a scrolling record of the tissue position as a function of time. Strong scatterers, such as organ boundaries, will be seen as bright spots on individual B-mode scan lines and horizontal streaks on the M-mode image if the boundary is stationary. Motion of these boundaries toward the transducer causes the streak to move upward on the display, and movement away causes the streak to move downward. This allows for easy measurement of movement. The M-mode has been used extensively to monitor heart-wall motion.

Doppler. The frequency shift caused by the reflection of sound from moving objects (acoustic Doppler effect) is used to image fluid flow in many application areas, including medicine. Blood flow in vessels can



(a) (b)

Fig. 10. Duplex display from a diagnostic ultrasound scanner. (*a*) B-mode image of the human carotid artery. The depth in centimeters is shown along the left side. The three triangles at left mark the position of the electronic focal zones for transmit. The graphic in the center of the image indicates the location where the Doppler measurement is being made. (*b*) Spectral Doppler display. The abscissa (horizontal axis) is time in seconds, and the ordinate (vertical axis) is flow velocity from the angle-corrected Doppler measurement. The gray scale in the display is the relative power in the Doppler signal for the velocity and time specified by the location on the display.

be imaged within the body in real time without the use of radioactive tracers. Commercially available devices can measure pulsatile flow velocities and can overlay frequency-shift (velocity) information, coded in color, on the traditional gray-scale B-scan image of the surrounding anatomy. Power-mode Doppler, in which the integrated Doppler spectral power is estimated and encoded in color, has been added to many ultrasound systems.

Duplex and triplex modes. The imaging information from a number of these modes can be combined in display formats termed duplex or triplex modes, in which ultrasound systems can continuously update the display at slow rates (10 images per second depending on image quality). **Figure 10** is an example of a duplex display, with both a B-mode image of the human carotid artery and the corresponding Doppler spectrum of the blood flow.

Nonlinear (harmonic) imaging. Nonlinear propagation is also being used in ultrasonic imaging. As an ultrasonic wave passes through tissue, higher harmonics will be produced in the portion of the field at high amplitude. Because of focusing, the higher amplitudes are confined to a narrower beam, and scattering from objects outside this beam is reduced. Therefore, emphasizing these nonlinear signals can produce better-resolution images. Artifacts introduced in the image from undesired reflections near the transducer, such as from ribs, are also reduced because the nonlinear components are produced only after a sufficient propagation distance. *See* BIOMEDICAL ULTRASONICS; MEDICAL IMAGING; MEDICAL ULTRASONIC TOMOGRAPHY; NONLINEAR ACOUSTICS.

Ultrasound contrast agents. Ultrasound contrast agents may provide additional means for overcoming signal-to-noise problems in smaller vessels. Most of these agents are stabilized gas bubbles, some of which are surrounded by shells of albumin, polysaccharides, or lipids. Some contain perfluorocarbon gases that have low solubilities and slow diffusion rates in aqueous media, which serve to slow their dissolution. Agents are used for imaging in the heart, brain, and other anatomic sites. Nonlinear imaging techniques are also being applied to these agents due to their strong nonlinear scattering that can produce substantially greater amplitudes in harmonics than tissue and even subharmonic emissions. In addition, higher-amplitude ultrasonic pulses can eliminate the signal from some contrast agents, and monitoring the rate at which signal returns provides a measure of blood flow.

Ultrasound bioeffects. The safety of medical diagnostic ultrasound has been studied almost since its initial use. In general, the approach has been to limit the acoustic output of ultrasonic devices to levels which have not produced observable bioeffects; that is, all diagnostic ultrasound devices should be equivalent in safety and effectiveness to pre-1976 ultrasound devices. Knowledge of potential thresholds for bioeffects has improved, and two indices have been identified to provide feedback to the user on potential bioeffects. The thermal index specifies the capability of raising the temperature of tissue. The mechanical index indicates the potential for the ultrasonic field to generate acoustic cavitation, which can produce thermal, chemical, and mechanical effects. These indices provide information for determining risk versus benefit when a physician determines that important information can be obtained at acoustic levels with bioeffects potential.

Ultrasonic therapy. Ultrasonic fields of sufficient amplitude can generate bioeffects in tissues. Although diagnostic ultrasound systems try to limit the potential for these effects, therapeutic levels of ultrasound have been used in medicine for a number of applications. Conventional therapeutic ultrasound is a commonly available technique used in physical therapy. High-frequency acoustic fields (typically 1 MHz) are applied through the skin to the affected area in either a continuous wave or long pulses.

Extracorporeal shock-wave lithotripsy (ESWL) disintegrates kidney stones with a high-amplitude acoustic pulse passing through the skin of the patient. The procedure eliminates the need for extensive surgery. Bioeffects are limited to the location of the stone by using highly focused fields which are targeted on the stone by imaging techniques such as ultrasound or fluoroscopy. The stone disruption process is not completely understood, but suggested mechanisms include shear stress and cavitation, which cause initial fracturing.

Ultrasound hyperthermia is used to treat a variety of conditions including cancer by heating the tissue to produce cell death. However, the technique is experimental in most applications. The hyperthermia systems typically use continuous-wave acoustic fields or acoustic pulses of considerably longer duration than those used in extracorporeal shock-wave lithotripsy.

Ultrasound ablative therapy uses higher-amplitude pulses to generate both mechanical and thermal disruption of the tissue.

Henry E. Bass; J. Brian Fowlkes; Veerle M. Keppens

Bibliography. C. E. Brennen, *Cavitation and Bubble Dynamics*, 1995; F. A. Duck, A. C. Baker, and H. C. Staritt, *Ultrasound in Medicine*, 1998; D. Ensminger, *Ultrasonics: The Low- and High-Intensity Applications*, 2d ed., 1988; D. H. Evans and W. N. McDicken, *Doppler Ultrasound: Physics, Instrumentation, and Signal Processing*, 2000; F. W. Kremkau, *Diagnostic Ultrasound: Principles and Instruments*, 1998; A. Migliori and J. Sarrao, *Resonant Ultrasound Spectroscopy*, 1997; E. P. Papadakis (ed.), *Ultrasonic Instruments and Devices*, 1999; K. S. Suslick, *Ultrasound: Its Chemical, Physical, and Biological Effects*, 1988.

Ultraviolet astronomy

Astronomical observations carried out in the region of the electromagnetic spectrum with wavelengths from approximately 10 to 350 nanometers. The ultraviolet spectrum is divided into the extreme-ultraviolet (EUV; 10–90 nm), far-ultraviolet (FUV;

90–200 nm), and near-ultraviolet (near-UV; 200–350 nm). Ultraviolet radiation from astronomical sources contains important diagnostic information about the composition and physical conditions of these objects. This information includes atomic absorption and emission lines of all the most abundant elements in many states of ionization. The hydrogen molecule (H_2), the most abundant molecule in the universe, has its absorption and emission lines in the far-ultraviolet. Thus, ultraviolet observations make it possible to probe a very wide range of physical conditions of matter in the universe, from the very cold gas in dense interstellar regions with temperatures of perhaps 30 K (−406°F) to the hot gas found in supernova remnants and in the coronas of stars and galaxies with temperatures approaching 10^6 K. *See* ASTRONOMICAL SPECTROSCOPY; ULTRAVIOLET RADIATION.

Observations. Ultraviolet radiation with wavelengths less than 310 nm is strongly absorbed by molecules in the atmosphere of the Earth. Therefore, ultraviolet observations must be carried out by using instrumentation situated above the atmosphere. Ultraviolet astronomy began with instrumentation at high altitudes aboard sounding rockets for brief glimpses of the Sun and stars. The first major ultraviolet satellite observatories to be placed in space were the United States *Orbiting Astronomical Observatories* (*OAO*s). *OAO 2* operated from 1968 to 1972 and provided the first full survey of the many kinds of ultraviolet sources in the sky, while *OAO 3* (*Copernicus*) operated from 1972 to 1980 and obtained high-resolution spectra of bright ultraviolet-emitting stars in order to probe the composition and physical state of intervening interstellar gas and to study the stellar winds of hot stars. Also, a number of smaller satellites, including the European *TD 1* and the Dutch *ANS*, provided very important survey measurements on the ultraviolet brightnesses of astronomical sources. *See* ROCKET ASTRONOMY.

With the launch of the *International Ultraviolet Explorer* (*IUE*) into a geosynchronous orbit in 1978, the full potential of ultraviolet astronomy to probe a wide range of scientific problems became a reality. The *IUE* satellite was a collaborative project of the U.S. National Aeronautics and Space Administration (NASA), the European Space Agency, and the United Kingdom Science and Engineering Research Council. It consisted of a reflecting telescope of modest size (18 in. or 45 cm in diameter) followed by several spectrographs with ultraviolet-sensitive detectors that produce ultraviolet spectra over the wavelength region from 120 to 320 nm. Between 1978 and 1996 the *IUE* obtained approximately 104,000 ultraviolet spectra of a wide range of astronomical objects, including comets and planets, cool and hot stars, exploding stars, external galaxies, and quasars.

In 1992 the extreme-ultraviolet window to the universe was opened with the launch of NASA's *Extreme Ultraviolet Explorer* (*EUVE*) satellite. The *EUVE* contains telescopes designed to produce images of the extreme-ultraviolet sky and spectra of bright extreme-ultraviolet sources in the wavelength range from approximately 10 to 90 nm. Because of strong absorption by neutral hydrogen in the interstellar gas, it was expected that the *EUVE* would not be able to see very many galactic sources. However, the irregular distribution of that gas has permitted the *EUVE* to probe to substantial distances in some directions. Most of the sources of radiation detected with the *EUVE* over $8^1/_2$ years of operation were stars with hot active outer atmospheres (or coronae) and hot white dwarf stars. *See* WHITE DWARF STAR.

Over the period 1985–1998, a number of missions were conducted or launched from the space shuttle, including the ASTRO missions containing the Hopkins Ultraviolet Telescope, the Ultraviolet Imaging Telescope, and the Wisconsin Ultraviolet Photo-Polarimeter Experiment. In addition, ultraviolet spectroscopy at medium and very high resolution has been accomplished with the Orbiting and Retrievable Far and Extreme Ultraviolet Spectrometer and the Interstellar Medium Absorption Profile Spectrograph. The scientific data for these missions were acquired over the 7–10-day period that the shuttle orbited the Earth.

Although it had an inauspicious beginning, the Hubble Space Telescope has been the centerpiece of both ultraviolet and visible astronomy. The telescope was launched into low Earth orbit in 1990, and shortly thereafter it was discovered that its 94-in.-diameter (2.4-m) primary mirror was ground incorrectly and the first images produced by the observatory were blurry. The problem was corrected in 1993 during a space shuttle refurbishment and repair mission. The complement of instruments aboard the Hubble Space Telescope has included imaging cameras operating at ultraviolet, visual, and infrared wavelengths, and spectrographs operating at ultraviolet and visual wavelengths. The two original Hubble spectrographs, the Faint Object Spectrograph and the Goddard High Resolution Spectrograph, were replaced with the Space Telescope Imaging Spectrograph in 1997. A very sensitive spectrograph designed to observe faint ultraviolet sources, the Cosmic Origins Spectrograph, was scheduled to be deployed on the Hubble Space Telescope in 2004 with the space shuttle *Columbia*. However, the *Columbia* atmospheric reentry disaster in 2003 led to the NASA decision in 2004 to cancel all future Hubble Space Telescope Shuttle servicing missions. It appears unlikely the Cosmic Origins Spectrograph will ever be deployed on the Hubble Space Telescope. With failing gyroscopes and batteries, the Hubble Space Telescope is not expected to be operational beyond 2008 or 2009 if there are no servicing missions. *See* HUBBLE SPACE TELESCOPE; SPECTROGRAPH; TELESCOPE.

The repaired Hubble Space Telescope has fulfilled its original expectations and provided spectacular high-angular-resolution pictures (better than 0.1 arcsecond) at visible and ultraviolet wavelengths, as well as low- and high-spectral-resolution ultraviolet spectra over the wavelength range from 120 to 320 nm. The ultraviolet measurements have been

used to pursue an extremely broad range of astronomical studies, from conditions in the atmospheres of the planets to measurements relevant to the creation of the elements in the first few minutes of the big bang. *See* BIG BANG THEORY.

The *Far-Ultraviolet Spectrograph Explorer* (*FUSE*) satellite was launched into an 800-km (500-mi) circular orbit in June 1999. *FUSE* is designed to explore the universe in the 90–120-nm region of the spectrum at high spectral resolution. The far-ultraviolet region of the spectrum is rich is diagnostic information about cold and hot gas in the interstellar and intergalactic medium and in the environments of stellar and extragalactic sources. The primary goals of the mission include measurements of the deuterium-to-hydgrogen abundance ratio in the Milky Way Galaxy and the intergalactic medium, and a study of the properties of hot plasmas and cold molecular hydrogen in the Galaxy. The deuterium abundance studies are providing insights about the production of deuterium in the big bang and the subsequent destruction of the element as it is cycled through stars.

Galaxy Evolution Explorer (*GALEX*), launched in 2003, is a modest-sized ultraviolet imaging and spectroscopic survey mission that will probe star formation over 80% of the age of the universe. The ultraviolet emissions detected by the far-ultraviolet detectors aboard *GALEX* are ideally suited for revealing the presence of hot young stars in external galaxies.

The *Cosmic Hot Interstellar Plasma Spectometer* (*CHIPS*), a small satellite launched in early 2003, is designed to carry out an all-sky study of the diffuse extreme-ultraviolet background radiation at wavelengths from 9 to 26 nm. *CHIPS* observations will be used to determine the temperature, ionization conditions, and cooling mechanisms of the hot gas believed to fill the region of space in the immediate vicinity of the Sun. The early results from this mission suggested the conditions in the local hot bubble are very different than expected.

Discoveries. The important discoveries of ultraviolet astronomy span all areas of modern astronomy and astrophysics. Some of the notable discoveries in the area of solar system astronomy include new information on the upper atmospheres of the planets, including planetary aurorae and the discovery of the enormous hydrogen halos surrounding comets. In studies of the interstellar medium, ultraviolet astronomy has provided fundamental information about the molecular hydrogen content of cold interstellar clouds along with the discovery of the hot phase of the interstellar medium, which is created by the supernova explosions of stars. In stellar astronomy, ultraviolet measurements led to important insights about the processes of mass loss through stellar winds and have permitted comprehensive studies of the conditions in the outer chromospheric and coronal layers of cool stars. The *IUE*, Hubble Space Telescope, and *FUSE* observatories have contributed to the understanding of the nature of the hot gaseous corona surrounding the Milky Way Galaxy. Ultraviolet observations of exotic astronomical objects, including exploding stars, active galactic nuclei, and quasars, have provided new insights about the physical processes affecting the behavior of matter in extreme environments. The spectrographs aboard the Hubble Space Telescope have revealed the existence of large numbers of hydrogen clouds in the intergalactic medium. These intergalactic clouds may contain much more normal (baryonic) matter than exists in the known luminous galaxies and stars. The measures of the abundance of deuterium in the Milky Way Galaxy and beyond have provided important constraints on the conditions in the evolving universe when it was only several minutes old. *See* COMET; COSMOLOGY; GALAXY, EXTERNAL; INTERSTELLAR MATTER; MILKY WAY GALAXY; PLANETARY PHYSICS; QUASAR; SATELLITE (ASTRONOMY); STAR; SUPERNOVA.

Blair D. Savage

Bibliography. J. N. Bahcall and L. Spitzer, Jr., The space telescope, *Sci. Amer.*, 247(1):40–51, July 1982; S. Bowyer, Extreme ultraviolet astronomy, *Sci. Amer.*, 271(2):32–39, August 1994; J. C. Brandt et al., The Goddard High Resolution Spectrograph: Instrument, goals, and science results, *Pub. Astron. Soc. Pacific*, 106:890–908, 1994; B. D. Savage, Ultraviolet-optical space astronomy: Past, present and future, in J. A. Morse, J. M. Shull, and A. L. Kinney (eds.), Ultraviolet-Optical Space Astronomy Beyond HST, *Astron. Soc. Pacific Conf. Ser.*, 164:3–14, 1999; B. D. Savage and K. R. Sembach, Interstellar abundances from absorption line observations with the Hubble Space Telescope, *Annu. Rev. Astron. Astrophys.*, 34:279–329, 1996; L. A. Shore, IUE: Nine years of astronomy, *Astronomy*, 15(4):14–22, April 1987.

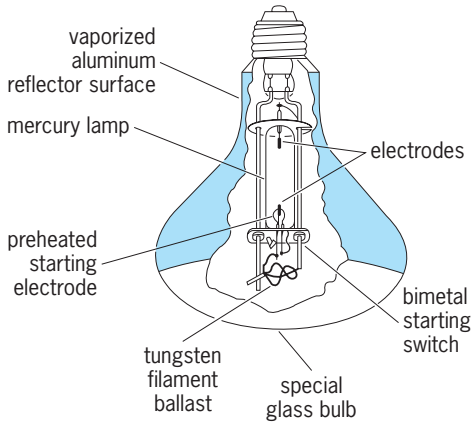
Ultraviolet lamp

A mercury-vapor lamp designed to produce ultraviolet radiation. Also, some fluorescent lamps and mercury-vapor lamps that produce light are used for ultraviolet effects. *See* FLUORESCENT LAMP; MERCURY-VAPOR LAMP; ULTRAVIOLET RADIATION.

Near-ultraviolet lamps. Ultraviolet energy in the wavelength region from 320 to 400 nanometers is known as near ultraviolet, or black light. Fluorescent and mercury lamps can be filtered so that visible energy is absorbed and emission is primarily in the black-light spectrum. The ultraviolet energy emitted is used to excite fluorescent pigments in paints, dyes, or natural materials to produce dramatic effects in advertising, decoration, and the theater; in industrial inspection, fluorescent effects are often used to detect flaws in machined parts and other products, as well as invisible laundry marks.

Middle-ultraviolet lamps. Middle ultraviolet spans the wavelength band from 280 to 320 nm. Mercury-vapor lamps are sometimes designed with pressures that produce maximum radiation in this region, using special glass bulbs that freely transmit this energy.

One such lamp type is the sunlamp. The **illustration** shows the reflector sunlamp, with a self-contained filament ballast and starting mechanism.



Cutaway view of ultraviolet lamp.

The reflector sunlamp combines the middle ultraviolet, which reddens the skin, with infrared energy and light from the filament to produce a suntanning effect with the sensations of warmth and brightness normally associated with sunshine. *See* SUNLAMP.

Other lamps designed for middle-ultraviolet radiation are known as photochemical lamps. They are used for a variety of tasks, including mold destruction, inspection of sheet metal for pinholes, and black-and-white printing of engineering drawings.

Far-ultraviolet lamps. Some radiation in the 220–280-nm wavelength band has the capacity to destroy certain kinds of bacteria. Mercury lamps designed to produce energy in this region (the 253.7-nm mercury line) are electrically identical with fluorescent lamps; they differ from fluorescent lamps in the absence of a phosphor coating and in the use of glass tubes that transmit far ultraviolet. Germicidal lamps are sometimes used to reduce airborne bacteria and to kill certain organisms on or near perishable products in storage or on certain products in the pharmaceutical industry. *See* ULTRAVIOLET RADIATION (BIOLOGY).

Alfred Makulec

Ultraviolet radiation

Electromagnetic radiation in the wavelength range 4–400 nanometers. The ultraviolet region begins at the short wavelength (violet) limit of visibility and extends to the wavelength of long x-rays. It is loosely divided into the near (400–300 nm), far (300– 200 nm), and extreme (below 200 nm) ultraviolet regions (see **illus.**). In the extreme ultraviolet, strong absorption of the radiation by air requires the use of evacuated apparatus; hence this region is called the vacuum ultraviolet. Important phenomena associated with ultraviolet radiation include biological effects and applications, the generation of fluorescence, and chemical analysis through characteristic absorption or fluorescence.

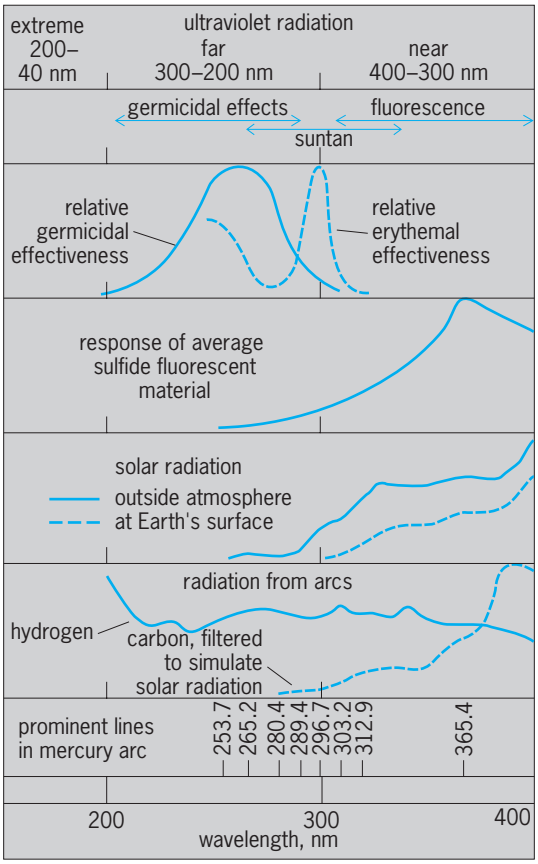
Biological effects of ultraviolet radiation include erythema or sunburn, pigmentation or tanning, and germicidal action. The wavelength regions responsible for these effects are indicated in the figure. Important biological uses of ultraviolet radiation include therapy, production of vitamin D, prevention and cure of rickets, and disinfection of air, water, and other substances. *See* ULTRAVIOLET RADIATION (BIOLOGY).

Fluorescence and phosphorescence are phenomena often generated as a result of the absorption of ultraviolet radiation. These phenomena are utilized in fluorescent lamps, in fluorescent dyes and pigments, in ultraviolet photography, and in phosphors. The effectiveness of ultraviolet radiation in generating fluorescence is shown in the illustration. *See* FLUORESCENCE; FLUORESCENT LAMP; PHOSPHORESCENCE; PHOTOGRAPHY.

Chemical analysis may be based on characteristic absorption of ultraviolet radiation. Alternatively, the fluorescence arising from absorption in the ultraviolet region may itself be analyzed or observed. *See* FLUORESCENCE MICROSCOPE; SPECTROSCOPY.

Sources of ultraviolet radiation include the Sun (although much solar ultraviolet radiation is absorbed in the atmosphere); arcs of elements such as carbon, hydrogen, and mercury; and incandescent bodies. The wavelengths produced by some sources of ultraviolet radiation are indicated in the illustration. *See* ULTRAVIOLET LAMP.

Artificial sources of ultraviolet light are often used to simulate the effects of solar ultraviolet radiation in the study of the deterioration of materials on exposure to sunlight. Trace amounts of chemicals which



Phenomena associated with ultraviolet radiation.

strongly absorb ultraviolet radiation may effectively stabilize materials against such degradation. *See* INHIBITOR (CHEMISTRY).

Detectors of ultraviolet radiation include biological and chemical systems (the skin, the eye of an infant, or eye without a lens, and photographic materials are sensitive to this radiation), but more useful are physical detectors such as phototubes, photovoltaic and photoconductive cells, and radiometric devices.

Fred W. Billmeyer

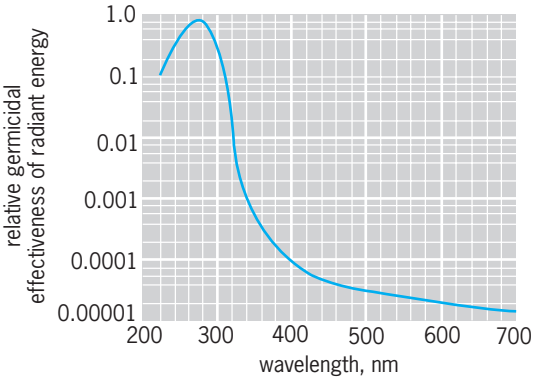
Bibliography. R. C. Denney and R. Sinclair, *Visible and Ultraviolet Spectroscopy*, 1988; T. Frost, *Ultraviolet Spectrometry*, 1992; W. Harm, *Biological Effects of Ultraviolet Radiation*, 1980; R. E. Huffman (ed.), *Ultraviolet Technology*, 3 vols., 1986, 1988, 1989; H. Moseley, *Nonionising Radiation: Microwaves, Ultraviolet Radiation and Lasers*, 1988.

Ultraviolet radiation (biology)

The ultraviolet portion of the spectrum includes all radiations from 15 to 390 nanometers (others extend this range to 4–400 nm). Radiations shorter than 200 nm are absorbed by most substances, even by air; therefore, they are technically difficult to use in biological experimentation. Radiations between 200 and 300 nm are selectively absorbed by organic matter, and produce the best-known effects of ultraviolet radiations in organisms. Radiations between 300 and 390 nm are relatively little absorbed and are less active on organisms. Ultraviolet radiations, in contrast to x-rays, do not penetrate far into larger organisms; therefore, the effects they produce are surface effects, such as sunburn and development of D vitamins from precursors present in skin or fur. The effects of ultraviolet radiations on life have, therefore, been assayed chiefly with unicellular organisms such as bacteria, yeast, and protozoa, although suspensions of cells of higher organisms, for example, eggs and blood corpuscles, have been useful as well.

Ultraviolet radiation in sunlight at the surface of the Earth is restricted to the span from about 287 to 390 nm, although shorter wavelengths are present beyond our atmosphere, as shown by measurements with rockets. Consequently, artificial sources of the radiations are generally used in experimentation. *See* ULTRAVIOLET RADIATION.

Photobiological effects. Only the ultraviolet radiations which are absorbed can produce photobiological action. All life activities are shown to be affected by ultraviolet radiations, the effect depending upon the dosage. Small dosages activate unfertilized eggs of marine animals, reduce the rate of cell division, decrease the synthesis of nucleic acid, especially in the nucleus, reduce the motility of cilia and of contractile vacuoles, and sensitize cells to heat. Large dosages increase the permeability of cells to various substances, inhibit most synthetic processes, produce mutations, stop division of cells, decrease the rate of respiration, and may even disrupt cells. The effect of ultraviolet radiations upon cells is invariably deleterious.



Relative bactericidal action of near-ultraviolet and visible regions for *Escherichia coli* on agar. (*After A. Hollaender, ed., Radiation Biology, vol. 2, McGraw-Hill, 1955*)

Despite their damaging effects, ultraviolet radiations are used as tools in biological research because they stop certain cell activities selectively without introduction of extraneous chemicals. They have been found especially useful in the production of mutations in microorganisms. *See* MUTATION.

Action spectra. Some wavelengths of ultraviolet radiations are more effective than others. More bacteria are killed by a given dosage of ultraviolet radiation at 260 nm than by the same dosage of radiation at 300 nm. When the bactericidal effectiveness of each of a series of wavelengths is plotted against the wavelength, the resulting curve is an action spectrum for the bactericidal effect (see **illus.**). Action spectra have been determined for many other effects of ultraviolet radiations. Each action spectrum is postulated to represent the absorption spectrum of the substance in the cell that is responsible for the particular effect. For the bactericidal effect, production of mutations, and retardation of cell division, the action spectrum suggests absorption by nucleoproteins of nucleic acid. For sensitization to heat and inhibition of ciliary movement, the action spectrum suggests absorption by ordinary proteins. For permeability, another action spectrum exists. *See* PHYSIOLOGICAL ACTION SPECTRA.

Mechanism. The effect of a given dosage of ultraviolet radiation upon protozoan cells is greater when the radiation is flashed than when continuous, that is, if a period of radiation is followed by a period of darkness. This indicates that a thermal reaction follows the primary photochemical reaction, a suggestion which is substantiated by the fact that increasing the temperature within the viable range accentuates the effect of flashing.

Action upon deoxyribonucleic acid (DNA), present in large amounts in chromosomes, consists primarily of the formation of pyrimidine dimers, chiefly between adjacent thymine residues on a strand. They interfere with replication of the DNA.

Photoreversal. The action of ultraviolet radiation on cells can be reversed to a considerable degree by simultaneous or subsequent exposure of the irradiated cells to short wavelength visible, violet and blue, or long wavelength ultraviolet light. This process

has been called photoreversal or photoreactivation. Thus, nucleic acid synthesis, inhibited by ultraviolet radiations, is resumed after exposure to visible light. At the same time, cell division, previously inhibited or retarded, is resumed. It appears that those effects of ultraviolet radiation having a nuclear site are most readily photoreversed.

Photoreversal consists of breaking of the thymine dimers into monomers, so reconstituting DNA in its original form. For this purpose a photoreactivating enzyme, which attaches itself to the dimers, is required as well as light.

Photoreversal is never complete; therefore, photoreactivated cells act as if they had been given a smaller dosage of ultraviolet radiations. It is evident that to make most effective use of ultraviolet radiations as a tool in experimental work, cells must be protected from visible light.

Dark repair. In addition to photoreactivation, it is now known that many cells kept from replication of DNA (in nonnutrient solutions) undergo dark repair. Pieces of DNA-containing dimers are excised (excision enzyme) and replaced on the basis of the information present in the other strand of DNA. Another enzyme (ligase) fastens the piece in place at the open end. Ultraviolet-resistant strains have very effective dark repair systems, while UV-sensitive ones do not.

Effects of ultraviolet on the skin. Erythema is the reddening of the skin following exposure to ultraviolet radiation of wavelength shorter than 320 nm, wavelength 296.7 nm being most effective. These radiations injure cells in the outer layer of the skin, or epidermis, liberating substances which diffuse to the inner layer of the skin, or dermis, causing enlargement of the small blood vessels. A minimal erythemal dose just induces reddening of the skin observed 10 h after exposure. A dose several times the minimal gives a sunburn, killing some cells in the epidermis after which serum and white blood cells accumulate, causing a blister. After the dried blister peels, the epidermis is temporarily thickened and pigment develops in the lower layers of the epidermis, both of these factors serving to protect against subsequent exposure to ultraviolet.

Both thickening of the epidermis and tanning may occur without blistering. Since the pigment in light-skinned races develops chiefly below the sensitive cells in the epidermis, it is not as effective as in dark-skinned races where the pigment is scattered throughout the epidermis. Consequently, the minimal erythemal dose is much higher for the dark- than for the light-skinned races.

Pigmentation or tanning also appears when the skin of young individuals is subjected to massive doses of ultraviolet radiations longer than 320 nm. Presumably this occurs by oxidation of precursors of the pigment, melanin, already present in the epidermis. Since such radiation is strong in sunlight, a skin may tan even in absence of short radiations.

Excessive exposure to ultraviolet radiation has been found to lead to cancer in mice, and it is claimed by some to cause cutaneous cancer in humans.

Clinical use. Ultraviolet radiations were once used extensively in the treatment of rickets, many skin diseases, tuberculosis other than pulmonary, especially skin tuberculosis (lupus vulgaris), and of many other diseases. The enthusiasm for sun bathing is, in part, a relic of the former importance of ultraviolet radiation as a clinical tool. Vitamin preparations, synthetic drugs, and antibiotics have either displaced ultraviolet radiations in such therapy or are used in conjunction with the radiations.

Ultraviolet radiations alone are still employed to treat rickets in individuals sensitive to vitamin D preparations. In conjunction with chemicals, they are used in treating skin diseases, for example, psoriasis, pityriasis rosea, and sometimes acne, as well as for the rare cases of sensitivity to visible light. They are also often used to sterilize air in hospitals. In some European laboratories, they are still used as adjuncts to drugs for treating lupus vulgaris and some other forms of tuberculosis. Ultraviolet radiations, however, are probably more important in research than in clinical practice. *See* RADIATION BIOLOGY; TUBERCULOSIS; VITAMIN D.

Arthur C. Giese

Bibliography. J. D. Longstreth (ed.), *Ultraviolet Radiation and Melanoma with a Special Focus on Assessing the Risks of Stratospheric Ozone Depletion*, 1987; H. Moseley, *Nonionizing Radiation: Microwaves, Ultraviolet and Laser Radiation*, 1988; J. A. Parrish (ed.), *The Effect of Ultraviolet Radiation on the Immune System*, 1983; M. Tevini (ed.), *UV-B Radiation: Effects on Humans, Animals, Plants, Microorganisms, and Materials*, 1993; World Health Organization Staff, *Solar and Ultraviolet Radiation*, 1992.

Ultrawideband (UWB) systems

An electronic system which has either an instantaneous bandwidth of at least 500 MHz or a fractional bandwidth (ratio of the instantaneous bandwidth measured at the −10 dB points to the center frequency) of at least 20%, whichever is greater. Although considered a recent breakthrough in broadband wireless technology, UWB has now experienced over 40 years of technology development.

Development of technology. Ultrawideband technology stems from work in time-domain electromagnetics begun in 1962 by Gerald F. Ross to describe fully the transient behavior of microwave networks through their characteristic impulse response. The conventional means of characterizing a linear, time-invariant system is by a swept, frequency response (that is, amplitude and phase measurements versus frequency), a very tedious endeavor for broadband systems. Ross, recognizing that a linear system could be fully characterized by its response to an impulse excitation (that is, the result of a single measurement), developed real-time hardware techniques to implement such a measurement. Mathematically, knowing the response to an impulse, the response to any arbitrary excitation could be readily determined by the convolution integral. However,

two fundamental problems existed: how to generate a close approximation to a mathematically ideal impulse (an infinitely large, infinitesimally short-duration waveform), and how to actually measure the extremely short duration responses that would be expected from such excitations.

In 1957, Leo Esaki invented the tunnel diode, the first known practical application of quantum physics. This device, with its extremely wide bandwidth (tens to hundreds of gigahertz), not only permitted subnanosecond pulse generation essential for impulse excitation but also could be used as a sensitive thresholding device for the subsequent detection of extremely short duration waveforms. In 1962, time-domain sampling oscilloscopes based on the tunnel diode were introduced for high-speed triggering and detection, first enabling the capture and display of ultrawideband waveforms. Impulse measurement techniques were subsequently applied to the analysis of wideband radiating antenna elements, in which the impulse response is the radiated electromagnetic field. In doing so, it quickly became apparent that short-pulse radar and even communications systems could be developed using the same set of tools. However, it was not until the introduction of a low-cost short-pulse receiver in 1972 to replace the expensive time-domain sampling oscilloscope that system developments in ultrawideband radar and communications accelerated rapidly. *See* ANTENNA (ELECTROMAGNETISM); TUNNEL DIODE; TUNNELING IN SOLIDS.

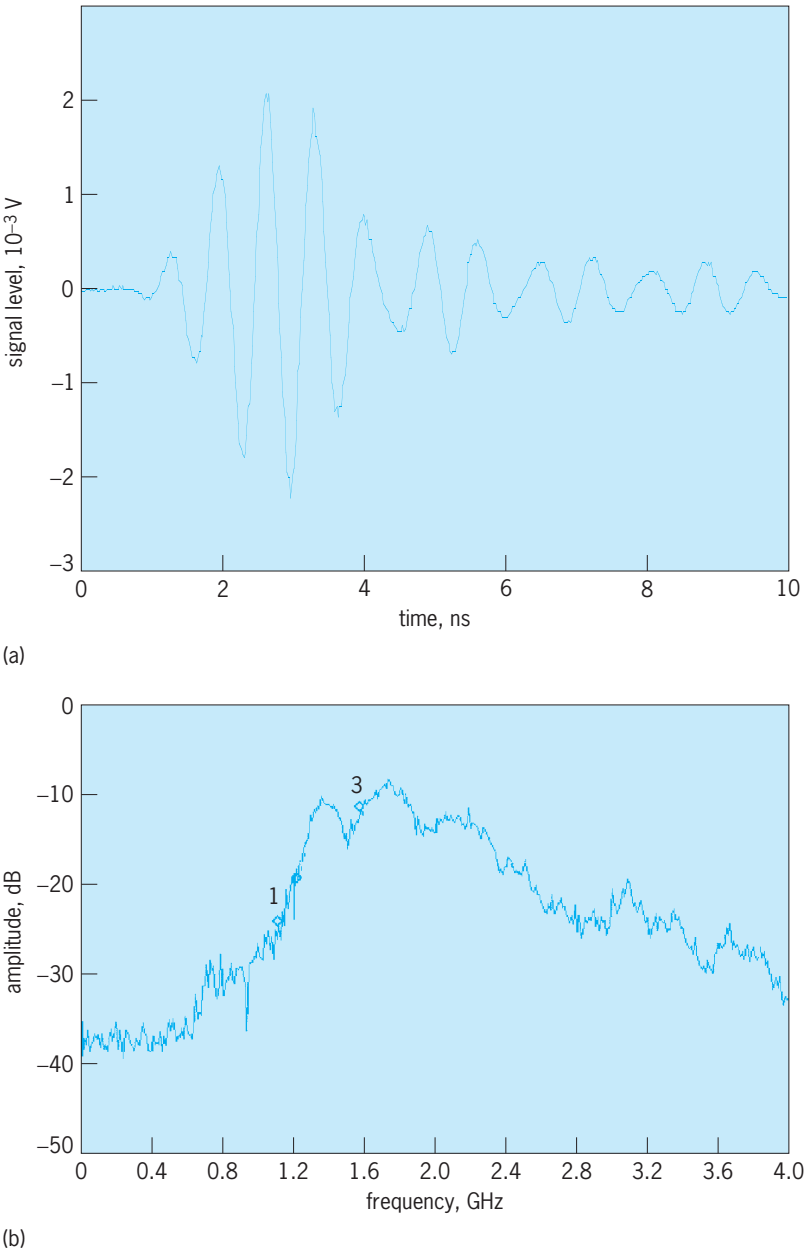


Fig. 1. Typical ultrawideband pulses in the time and frequency domains. (*a*) Typical ultrawideband transmit pulse in free space, with just a few cycles of radio-frequency energy. (*b*) Typical UWB power spectrum from impulse-excited wideband antenna. Center frequency $f_0 = 1.78$ GHz, and bandwidth at −10 dB relative to peak amplitude is 1.12 GHz, giving a fractional bandwidth of 63%.

Early techniques for ultrawideband signal generation used the fast rise (or fall) times of a baseband-generated pulse to impulse- or shock-excite a wideband antenna, which in turn would generate an electromagnetic burst consisting of only a few radio-frequency cycles of energy (**Fig. 1**). Tunnel, step-recovery, and avalanche diodes were used to produce this excitation signal, while modern techniques use the fast rise time of a semiconductor switch or gate. By varying the dimensions of the antenna, the frequency and bandwidth characteristics of the resulting ultrawideband pulse could be adjusted. Modern techniques for ultrawideband pulse generation use time-domain filtering techniques for spectral shaping.

Until the late 1980s, ultrawideband technology was alternatively referred to as baseband, carrier-free, or impulse. The term "ultrawideband" was not applied until approximately 1989, when the theory, techniques, and many hardware approaches of ultrawideband had experienced nearly 30 years of development. Somewhat paralleling the development of spread-spectrum systems after World War II, much of the early (pre-1994) development of ultrawideband systems (particularly for communications applications) was classified. The fact that these extremely short duration pulses required unique techniques for detection made them of particular interest for low-probability-of-detection applications for the government and military.

Significance of technology. The early interest in short-pulse techniques was primarily concentrated in short-range radar applications. From first principles, the shorter a radar's pulse duration, the finer the resultant range resolution. Thus, with pulse durations measured in hundreds of picoseconds, resolutions of only a few inches (several centimeters) or less were now possible. Such resolutions were orders of magnitude better than what could be achieved at that time, and significant research in the use of short-pulse techniques for radar applications continued.

From the time-scaling property of the Fourier transform relationship between time and frequency domains, the shorter the transmitted pulse, the wider the instantaneous bandwidth. Thus, since short-pulse waveforms could be inexpensively produced

directly at baseband (that is, without modulating a carrier frequency as in conventional radar), applications of ultrawideband to ground-penetrating radar and through-wall imaging followed quickly. Here, low-frequency wideband excitation translated into significantly enhanced material penetration capability. Ground-penetrating radar is used to locate buried objects, including humans in disasters such as building or mine collapses. Through-wall imaging is used for law enforcement and intelligence activities. *See* FOURIER SERIES AND TRANSFORMS; GROUND-PROBING RADAR; RADAR.

The early interest in ultrawideband for communications stemmed from the fact that short-pulse waveforms are extremely difficult to detect. Since the pulse bandwidth is spread over many hundreds of megahertz to gigahertz, a conventional (that is, narrowband) receiver will receive only a small slice of this spectral energy. The total received energy is directly proportional to the ratio of the intercept receiver to ultrawideband spread bandwidths.

These short-pulse waveforms have other unique properties as well. For example, one limiting factor affecting the performance of both mobile and indoor communications systems, particularly with narrowband systems, is the deleterious effect of multipath signal cancellation. In general, a received signal is the composite sum of the direct (or line-of-sight) signal from transmitter to receiver and a number of secondary reflections of the signal off objects between the transmitter and receiver. These "bounced" or multipath returns will arrive at the receiver later than the direct signal and, because of this time misalignment and possible phase inversion due to the reflection, will cause varying amounts of signal cancellation or distortion of the direct path. However, if the duration of the pulse is short enough, it is possible to distinguish in time between the direct and reflected multipath returns without causing any signal cancellation.

This capability is illustrated by the geometry of **Fig. 2**, in which two units, A and B, are attempting to communicate, but the signal also bounces off a wall located 3 m (10 ft) away. With the speed of light at roughly 0.3 m (1 ft) per nanosecond, a signal reflected from the wall will arrive approximately 10 ns (3-m or 10-ft path differential) later than the direct path. If the signal's pulse duration is less than this 10-ns differential, the direct and reflected path can be fully resolved without mutual interference. Thus, short-pulse techniques become of particular interest for high-reliability communications in severe multipath channels. However, if the delay spread of the channel (that is, the time over which such multipath effects are dominant) exceeds the bit duration (reciprocal bit rate) of the ultrawideband communications system, intersymbol interference (distortion) can occur and ultrawideband techniques become less effective.

Finally, because the time of arrival of a short-duration pulse can be very accurately measured using high-speed detection circuitry, such as enabled by the Esaki diode, accurate time-of-flight or distance measurements can he accomplished, opening up the use of the technology for precision localization applications.

Applications. Since the U.S. Federal Communications Commission (FCC) first established rules for the unlicensed use of ultrawideband technology in 2002, only a handful of ultrawideband designs have progressed beyond the laboratory stage.

Much of the short-pulse ultrawideband work either remains classified or has been used for applications having primarily government or military interest. Recent applications include an ultrawideband airborne wireless intercommunications system (AWICS) which uses the low-probability-of-intercept and multipath mitigation features of the ultrawideband waveform to provide a wireless communications capability within an aircraft or helicopter fuselage; ultrawideband wireless transceivers for the relay of voice, data and compressed video for covert network communications; ultrawideband perimeter intrusion radars for detecting unauthorized personnel; ultrawideband through-wall sensing radars for police and fire rescue; ultrawideband collision and obstacle avoidance radars for unmanned aerial vehicle (UAV) guidance and control; and several others.

On the commercial side, much of the lag in getting ultrawideband products to market has been because of the need for consensus on industry standards. The Institute of Electrical and Electronic Engineers (IEEE) is sponsoring two industry-wide standards for ultrawideband: IEEE 802.15.3a dealing with high-speed ultrawideband systems for wireless personal area networks (WPAN), and IEEE 802.15.4a dealing with low-speed ultrawideband systems for application to longer-range wireless networking and localization (for example, tracking of ultrawideband tags).

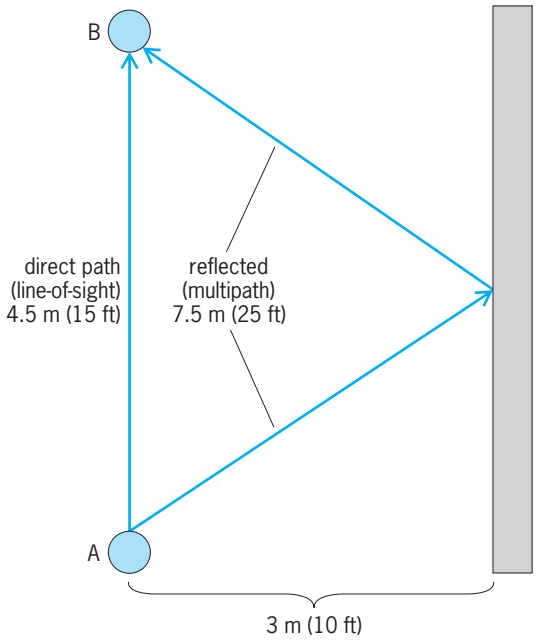


Fig. 2. Multipath geometry.

The high-speed WPAN standards group is split into two major camps, with one favoring the use of direct-sequence spread spectrum (DSSS) techniques and the other favoring the use of multiband orthogonal frequency division multiplex (M-OFDM). In both, the modulations chosen represent ultrawideband variants of more conventional waveforms. That is, instead of using short-pulse techniques to generate a large instantaneous bandwidth, these ultrawideband variants rely upon the underlying high-speed data stream to produce the requisite bandwidth. For example, DSSS techniques are used in IEEE 802.11b, cellular wireless, and various wireless home products (such as wireless telephones), while OFDM techniques are commonly found in asymmetric digital subscriber line (ASDL) services, IEEE 802.11a/g (WiFi) systems, digital audio broadcast (DAB), and digital terrestrial television broadcast in Europe and Japan. It appears that instead of a single standard for WPAN, nonstandard consumer products based on both of these technologies will emerge. *See* SPREAD SPECTRUM COMMUNICATION.

The low-speed ultrawideband localization standards group IEEE 802.15.4a in April 2005 completed the first step of soliciting proposals from industry, with the goal of reaching a consensus in the future. However, it appears that the vast majority in industry favor a physical layer (PHY) for ultrawideband localization, based on the use of short-pulse techniques. Unlike WPAN, several companies already have ultrawideband localization products on the market. According to measured data from these systems, the use of short-pulse techniques, together with precise time-of-flight measurements, enables 2- and 3D localization to within 15 cm (6 in.), even in the presence of severe multipath as is typically found in hospitals and industrial plants.

FCC report and order. As with any radio-frequency (RF) technology, both commercial and government use of ultrawideband requires appropriate frequency allocations. On February 14, 2002, the FCC approved a First Report and Order (R&O) to permit ultrawideband technology under Part 15 of its regulations. The R&O was precedent-setting in that, for the first time, the FCC had permitted intentional emissions to fall into previously restricted areas of the spectrum. The FCC waved restrictions on ultrawideband operation within certain restricted bands, but imposed additional constraints on the use of the technology for imaging, radar, and communications applications. These constraints included limitations on the field of operation and requirements for interagency coordination for the use of imaging and through-wall sensors, additional peak power constraints for all forms of ultrawideband operation, and operational restrictions on peer-to-peer indoor and handheld outdoor ultrawideband communications devices. Under this First R&O, ultrawideband communications devices are restricted to intentional operation between 3.1 and 14.6 GHz; through-wall imaging and surveillance systems are restricted to between 1.99 and 10.6 GHz (for use only by law enforcement, fire and rescue, and other designated organizations); and automotive radars are restricted to frequencies above 24.075 GHz. Ground-penetrating radar and through-wall imaging sensors were also permitted below 960 MHz.

On December 15, 2004, the FCC released a Second R&O on ultrawideband technology which opened up a broad range of new applications for both ultrawideband and wideband systems. Recognizing that modern ultrawideband devices can be spectrally confined to nonrestricted frequency bands, the FCC created two new Part 15 sections: 15.250 and 15.252.

Section 15.250 permits the operation of wideband systems (that is, systems now defined with a minimum bandwidth of at least 50 MHz) within the band 5925–7250 MHz. Except for operation on an aircraft or satellite or within a toy, wideband or ultrawideband systems operating under 15.250 have no further restrictions on their use as was previously mandated under the First R&O. Thus, indoor/outdoor, vehicular and shipboard, and other uses are now permitted at power levels equivalent to those specified in the First R&O.

The new Section 15.252 opens up two new frequency bands for the operation of wideband vehicular radar systems: 16.2–17.7 GHz and 23.12–29.0 GHz. The technology used in these frequency bands is envisioned for field-disturbance sensors providing automotive back-up assistance.

In January 2005, the United Kingdom's Office of Communications (Ofcom) released a consultation document that proposed rules for the use of ultrawideband technology within the U.K. Comments have been received from over 60 international government and corporate entities, and it is expected that the U.K. will incorporate rules for the unlicensed use of ultrawideband. However, it is anticipated that the U.K. will coordinate its rulemaking with the final recommendations from the International Telecommunication Union (ITU) Radiocommunication Bureau's Task Group 1/8 when it completes its study on the compatibility between ultrawideband devices and radiocommunication services. An ITU decision, representing the 189 member states, will have global ramifications for the introduction of ultrawideband technology.

Robert J. Fontana

Bibliography. C. L. Bennett and G. Ross, Time-domain electromagnetics and its applications, *Proc. IEEE*, 66(3):299–318, 1978; J. Foerster et al., Ultrawideband technology for short- or medium-range wireless communications, *Intel Technol. J.*, 2d Quarter, 2001; D. G. Leeper, Wireless data blaster, *Sci. Amer.*, May 2002; J. Taylor (ed.), *Introduction to Ultra-Wideband Radar Systems*, CRC Press, Boca Raton, FL, 1995; *2002 IEEE Conference on Ultra Wideband Systems and Technologies*, Institute of Electrical and Electronic Engineers, May 21–23, 2002; *2003 IEEE Conference on Ultra Wideband Systems and Technologies*, Institute of Electrical and Electronic Engineers, Nov. 16–19, 2003; *Ultra Wideband Systems 2004: Joint Conference on Ultra Wideband Systems and Technologies and International Workshop on UWB Systems*, Institute of Electrical and Electronics Engineers, May 18–21, 2004;

Ultra-Wideband, Short-Pulse Electromagnetics 1, 2, 3, and *4*, Plenum Press, New York, 1993, 1994, 1997, 1999.

Umklapp process

A concept in the theory of transport properties of solids which has to do with the interaction of three or more waves, such as lattice waves or electron waves, in a solid. In a continuum, such interactions occur only among waves described by wave vectors $\mathbf{k}_1$, $\mathbf{k}_2$, and so on, such that the interference condition, given by Eq. (1), is satisfied. The sign of $\mathbf{k}$

$$\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3 = 0 \tag{1}$$

depends on whether the wave absorbs or emits energy. Since $\hbar\mathbf{k}$ is the momentum of a quantum (or particle) described by the wave, Eq. (1) corresponds to conservation of momentum. In a crystal lattice further interactions occur, satisfying Eq. (2), where $\mathbf{b}$ is

$$\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3 = \mathbf{b} \tag{2}$$

any integral combination of the three inverse lattice vectors $\mathbf{b}_i$, defined by $\mathbf{a} \cdot \mathbf{b}_j = 2\pi\delta_{ij}$, the $\mathbf{a}$'s being the periodicity vectors. The processes described by Eq. (2) are Umklapp processes or flip-over processes, so called because the total momentum of the initial particles or quanta is reversed. *See* CRYSTAL STRUCTURE.

Examples of Umklapp processes are the following: (1) interactions of three lattice waves due to anharmonic lattice forces; of these, only processes described by Eq. (2) produce intrinsic thermal resistance in nonmetals; the exponential variation of the thermal resistance observed in dielectric crystals at low temperatures confirms the concept of Umklapp processes; (2) scattering of electrons by lattice waves, causing electrical and thermal resistance in metals; it has become clear that the observed properties cannot be accounted for in terms of processes described only by Eq. (1), but Umklapp processes must also be considered; (3) Bragg reflection, which can be regarded as an Umklapp process involving only two waves. *See* CONDUCTION (HEAT); ELECTRICAL CONDUCTIVITY OF METALS; THERMAL CONDUCTION IN SOLIDS; X-RAY DIFFRACTION. Paul G. Klemens

Bibliography. C. Kittel, *Introduction to Solid State Physics*, 7th ed., 1996; J. M. Ziman, *Principles of the Theory of Solids*, 2d ed., 1979.

Uncertainty principle

A fundamental principle of quantum mechanics, which asserts that it is not possible to know both the position and momentum of an object with arbitrary accuracy. This contrasts with classical physics, where the position and momentum of an object can both be known exactly. In quantum mechanics, this is no longer possible, even in principle. More precisely, the indeterminacy or uncertainty principle, derived by W. Heisenberg, asserts that the product of Δx and Δp—measures of indeterminacy of a coordinate and of momentum along that coordinate—must satisfy inequality (1). The Planck constant,

$$\Delta x \times \Delta p \gtrsim \hbar = \frac{h}{2\pi} \tag{1}$$

$h \simeq 6.6310^{-34}$ joule-second, is very small, which makes inequality (1) unimportant for the measurements that are carried out in everyday life. Nevertheless, the consequences of the inequality are critically important for the interactions between the elementary constituents of matter, and are reflected in many of the properties of matter that are ordinarily taken for granted.

Indeterminacy and atomic sizes. For example, the density of solids and liquids is set to a large degree by the uncertainty principle, because the sizes of atoms are determined with decisive help of inequality (1). In the ground state of the hydrogen atom, an electron is bound to a proton with an energy E_B equal to the sum of the potential energy (which is $-e^2/a$, because of the Coulomb attraction of two particles of charge e and $-e$, separated by the distance a) and the kinetic energy, as given in Eq. (2), where p and m_e

$$E_B = \frac{-e^2}{a} + \frac{p^2}{2m_e} \tag{2}$$

are the momentum of the electron with respect to the nucleus and its mass. In the absence of quantum indeterminacy, both p and a could be simultaneously arbitrarily small, and E_B could be arbitrarily negative. However, in quantum theory inequality (1) implies that the smallest momentum for an electron localized to within the distance a of a stationary proton is no less than $p \simeq \hbar/a$. Substituting this relation in Eq. (2) and minimizing E_B with respect to a leads to the conclusion that the radius of the hydrogen atom in the ground state is given by Eq. (3). This is

$$a_0 = \frac{h^2}{4\pi^2 m_e e^2} \simeq 0.529 \times 10^{-10} \text{ meter} \tag{3}$$

known as a radius of the first Bohr orbit of the hydrogen atom. The corresponding binding energy $E_B = -e^2/2a_0 \simeq -13.6$ eV, in agreement with much more elaborate calculations and with the measurements. *See* ATOMIC STRUCTURE AND SPECTRA.

When the atoms are tightly packed (separated by approximately $2a_0$), the typical density of the resulting material is equal to the proton mass divided by the volume of a cube of size $2a_0$, approximately 1.4 g/cm^3. While this calculation is a rough order-of-magnitude estimate, as it ignores complexities of atomic structure and the presence of neutrons in the nuclei (which would approximately double the density), and does not take into account the intricacies of atomic packing, it yields an essentially correct guess about the typical densities of solids and liquids. *See* QUANTUM THEORY OF MATTER.

Indeterminacy and the wave function. The wave-particle duality suggests a revealing perspective on quantum indeterminacy. States of quantum objects

can be represented by a wave function $\psi(x)$, which evolves in accord with the linear Schrödinger equation. Wave functions of definite wavelength λ (or wave number $k = 2\pi/\lambda$) are given by Eq. (4), and

$$\psi(x) \sim \exp(-ikx) \tag{4}$$

describe a free particle with a definite momentum $p = \hbar k = h/\lambda$. Almost any function of x (and, in particular, any wave function) can be expressed as a sum of such special plane-wave (definite-momentum) solutions, as in Eq. (5*a*). Here $\tilde{\psi}(k)$ supplies the complex weights of the different plane waves, and is known as a Fourier transform of $\psi(x)$. It can be obtained through Eq. (5*b*). The Fourier transform is

$$\psi(x) = (2\pi)^{-1/2} \int dk \tilde{\psi}(k) \exp(-ikx) \tag{5a}$$

$$\tilde{\psi}(k) = (2\pi)^{-1/2} \int dx \psi(x) \exp(ikx) \tag{5b}$$

an alternative, complete description of the same abstract state of the quantum, but in terms of the momentum rather than position: Both $\psi(x)$ and $\tilde{\psi}(k)$ contain exactly the same information. Thus, it is not possible to alter either of them [by, say, confining $\psi(x)$ to a smaller interval Δx] without affecting its transform. *See* SCHRÖDINGER'S WAVE EQUATION; WAVE MECHANICS.

It is a mathematical fact derived from the properties of Fourier transforms that a wave function $\psi(x)$ which is peaked within $\Delta(x)$ in position corresponds to $\tilde{\psi}(k)$ with a range of wave numbers, Δk, equal to or greater than approximately $1/\Delta x$ (**Fig. 1**). This leads to an equivalent statement of the uncertainty principle, Eq. (6). It turns out that the absolute min-

$$\Delta x \times \Delta k \gtrsim 1 \tag{6}$$

imum of this product is achieved for gaussian wave

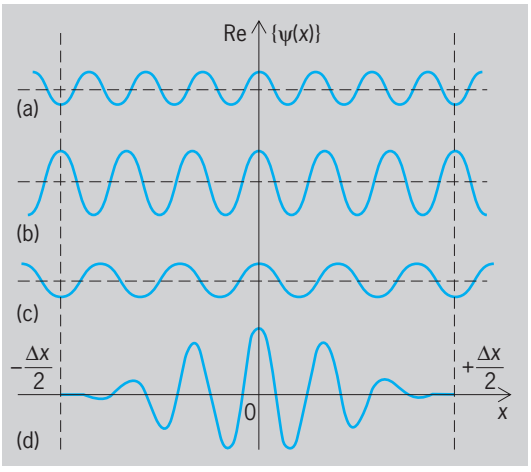


Fig. 1. Real parts of three wave functions with definite wave numbers (*a*) $k_0 + (\Delta k/2)$, (*b*) k_0, and (*c*) $k_0 - (\Delta k/2)$; and (*d*) their sum. The three wave functions add up constructively at $x = 0$, but interfere destructively at $+\Delta x/2$ and $-\Delta x/2$, resulting in a localized wave packet with a half-width inversely proportional to Δk. (*After C. Cohen-Tannoudji, B. Diu, and F. Laloë, Quantum Mechanics, Wiley, Hermann, 1977*)

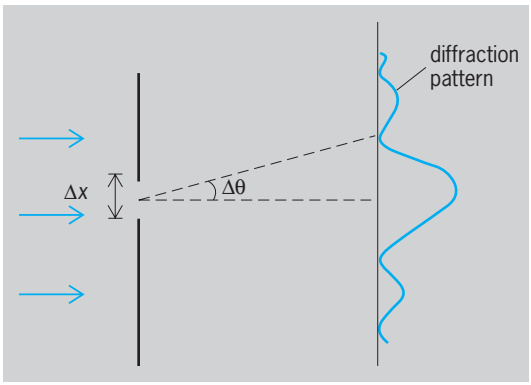


Fig. 2. Indeterminacy and a single-slit diffraction pattern. The size of the slit, Δx, is a measure of the uncertainty in position of the quantum. As the slit becomes narrower, the momentum in the direction parallel to the screen with the slit becomes more uncertain, resulting in a larger deflection angle, $\Delta\theta$, and a broader diffraction pattern. (*After R. P. Feynman, R. B. Leighton, and M. Sands, Feynman Lectures on Physics, vol. 1, Addison–Wesley, 1964, reprint 1989*)

functions. These are given by Eq. (7*a*), where σ is a parameter, and their corresponding Fourier transforms by Eq. (7*b*). The ranges Δx and Δk can be

$$\psi(x) \sim \exp(-x^2/2\sigma^2) \tag{7a}$$

$$\tilde{\psi}(k) \sim \exp(-k^2\sigma^2/2) \tag{7b}$$

defined as second moments of the probability distributions corresponding to $\psi(k)$ and $\tilde{\psi}(k)$. For the two gaussians in Eqs. (7), this yields $\Delta x = \sigma/\sqrt{2}$ and $\Delta k = 1/\sqrt{2}\sigma$. Hence, when the uncertainties are defined as dispersions, the minimum of the product of position and momentum is given by Eq. (8). The gaus-

$$\Delta x \Delta p = \frac{\hbar}{2} \tag{8}$$

sian states which attain this minimum of uncertainty are sometimes called coherent states, and can be realized as a ground state of a harmonic oscillator. *See* COHERENCE; DISTRIBUTION (PROBABILITY); FOURIER SERIES AND TRANSFORMS; HARMONIC OSCILLATOR.

Indeterminacy and predictability. In classical physics, simultaneous knowledge of position and momentum can be used to predict the future trajectory of a particle. Quantum indeterminacy and the limitations it imposes force such classical notions of causality to be abandoned. This can be illustrated in a single-slit diffraction pattern (**Fig. 2**). Quanta arrive one by one from the left, each starting with a well-defined and identical momentum. They can pass through a screen with an opening (size Δx), so that to the right of the slit the position of each quantum in x is known with the accuracy determined by the size of the opening. However, as the wave packet is now restricted in x, its momentum along the x axis must be uncertain. As a result, when the slit becomes narrower, the diffraction pattern forming on a plate to the right of the screen (which can be thought of as a photographic plate) will become broader. Quanta will arrive one by one, blackening the photographic plate one grain at a time, but the pattern of many such

quanta will have a distinctly wavelike character. This further illustrates the wave-particle duality, characteristic of quantum theory. It also shows that prediction in quantum mechanics is statistical in character. In the example at hand, it is expressed by the probability density given by Eq. (9), which is evaluated at

$$p(x) = |\psi(x)|^2 \tag{9}$$

the point of detection, along the photographic plate.

Other uncertainty relations. Another well-known example of indeterminacy involves energy and time, as given by inequality (10). Physically, its origins are

$$\Delta E \times \Delta t \gtrsim \hbar \tag{10}$$

somewhat different from those of inequality (1), but mathematically this inequality again derives from the properties of Fourier transforms. Inequality (10) relates, for example, lifetimes of unstable states with the widths of their lines. *See* LINEWIDTH.

In quantum physics, relations similar to inequalities (1) and (10) hold for pairs of many other quantities. They demonstrate that the acquisition of the information about a quantum object cannot be usually achieved without altering its state. Much of the strangeness of quantum physics can be traced to this impossibility of separating the information about the state from the state itself. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS.

Wojciech Hubert Zurek

Bibliography. C. Cohen-Tannoudji, B. Diu, and F. Laloë, *Quantum Mechanics*, 1977; R. P. Feynman, R. B. Leighton, and M. Sands, *Feynman Lectures on Physics*, 1964, reprint 1989; J. A. Wheeler and W. H. Zurek, *Quantum Theory and Measurement*, 1983.

Unconformity

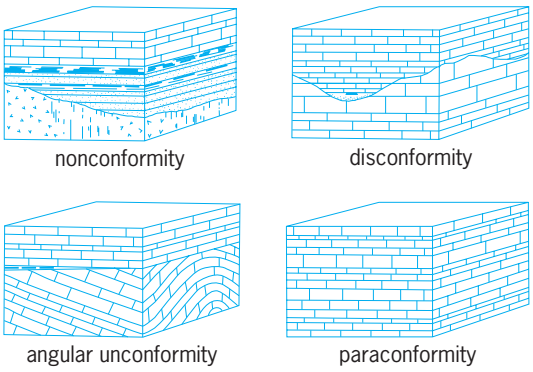
In the stratigraphic sequence of the Earth's crust, a surface of erosion that cuts the underlying rocks and is overlain by sedimentary strata. The unconformity represents an interval of geologic time, called the hiatus, during which no deposition occurred and erosion removed preexisting rock. The result is a gap, in some cases encompassing millions of years, in the stratigraphic record.

Types. There are four kinds of unconformable relations (see **illus.**):

1. Nonconformity—underlying rocks are not stratified, such as massive crystalline rocks formed deep in the Earth.

2. Angular unconformity—underlying rocks are stratified but were deformed before being eroded, resulting in angular discordance; this was the first type to be recognized; the term unconformity was originally used to describe the geometric relationship between the underlying and overlying bedding planes.

3. Disconformity—underlying strata are undeformed and parallel to overlying strata, but separated by an evident erosion surface.



Four types of unconformity. (*After C.O. Dunbar and J. Rodgers, Principles of Stratigraphy, John Wiley and Sons, 1957*)

4. Paraconformity—strata are parallel and the erosion surface is subtle, or even indistinguishable from a simple bedding plane.

Importance in geologic record. Nonconformities imply deep or long-continued erosion. Angular unconformities are formed during episodes of mountain building; strata are deformed and then uplifted into the zone of erosion. Disconformities can be formed either by broad uplift (without deformation) or by a drop in sea level. Sea-level falls produce extensive disconformities that serve to define sequences of intervening strata. Several large-scale falls have generated continent-wide erosion surfaces; more frequent smaller-scale falls have affected only the continental margins. Marginal sequences can be recognized in the subsurface by using geophysical seismic profiling. Paraconformities are produced in the same fashion as disconformities, but their recognition is based on gaps in the biostratigraphic record; missing fossil zones indicate the breaks in deposition. On the deep ocean floor, paraconformities are formed both by physical erosion and by chemical dissolution of the sediment. The breaks correspond to climatic and tectonic changes that control the formation and circulation of oceanic bottom water. *See* STRATIGRAPHY.

Charles W. Byers

Underground mining

The extraction of ore from beneath the surface of the ground. Underground mining is also applied to deposits of industrial (nonmetallic) minerals and rocks, and underground or deep methods are used in coal mining. Some ores and industrial minerals can be recovered from beneath the ground surface by solution mining or in-place leaching using boreholes. *See* COAL MINING; SOLUTION MINING.

Underground mining involves a larger capital investment and higher production cost per ton of ore than open pit mining. It is done where mineral deposits are situated beyond the economic depth of open pit mining; it is generally applied to steeply dipping or thin deposits and to disseminated or massive deposits for which the cost of removing the overburden and the maintaining of a slope angle in

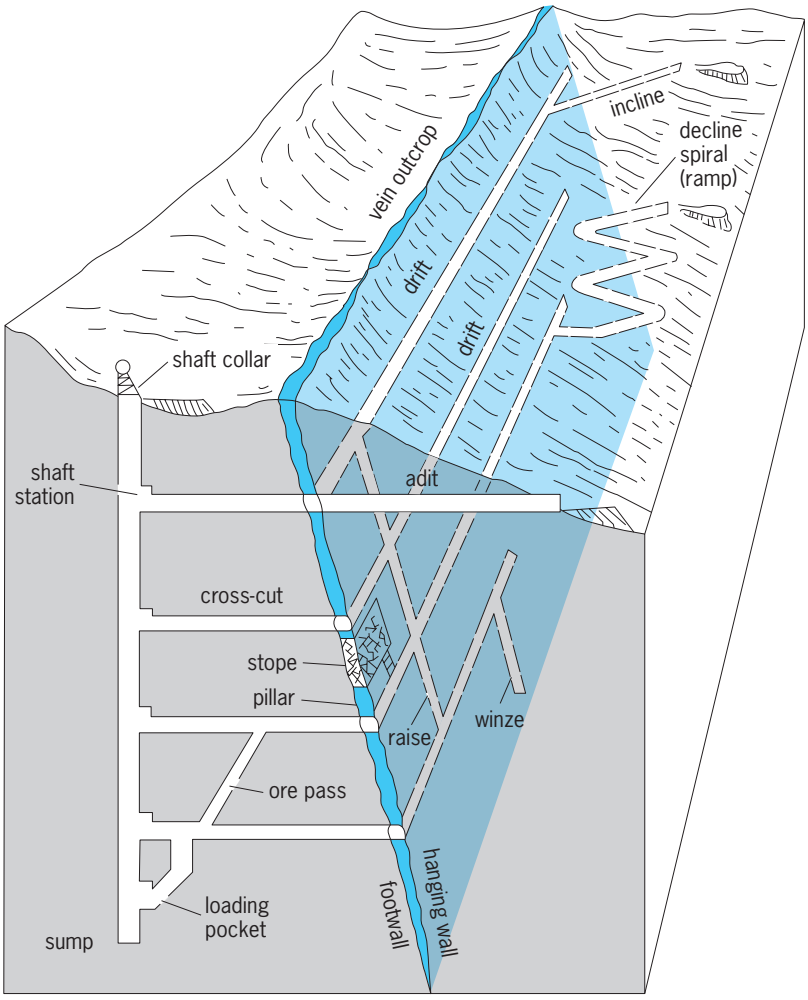


Fig. 1. Underground mining entries and workings.

adjacent waste rock would be prohibitive. In some situations, the shallower portion of a large orebody will be mined by open pit methods, and the deeper portion will be mined by underground methods. *See* OPEN PIT MINING.

Underground mine entries are by shaft, adit, incline, or spiral ramp (**Fig. 1**). Development workings, passageways for gaining access to the orebody from stations on individual mine levels, are called drifts if they follow the trend of the mineralization, and cross-cuts if they are driven across the mineralization. Workings on successive mine levels are connected by raises, passageways that are driven upward. Winzes are passageways that are sunk downward, generally from a lowermost mine level.

In a fully developed mine with a network of levels, sublevels, and raises for access, haulage, pumping, and ventilation, the ore is mined from excavations referred to as stopes. Pillars of unmined material are left between stopes and other workings for temporary or permanent natural support. In large-scale mining methods and in methods where an orebody and its overlying waste rock are allowed to break and cave under their own weight, the ore is extracted in large collective units called blocks, panels, or slices. *See* MINING.

Exploration

Exploration and development constitute the preproduction stage of underground mining. Exploration refers to the delineation of a newly discovered mineral deposit or an extension of a known deposit and to its evaluation as a prospect. During exploration, the deposit is investigated in sufficient detail to estimate its tonnage and grade, its metallurgical recovery characteristics, and its suitability for mining by various methods.

Information on the size, shape, and attitude of a deposit and information for estimating the tonnage and grade of the ore is taken from drill holes and underground exploration workings. Diamond core drilling provides intact samples of ore and rock for assaying and for detailed geologic and geotechnical study; percussion drilling provides chips of material for the recognition of ore and waste boundaries and for additional sampling. Underground exploration workings are used for bulk and detailed sampling, rock mechanics testing, and the siting of machinery for underground drilling. *See* DRILLING AND BORING, GEOTECHNICAL.

The tonnage and the grade of the material available in a mineral deposit are interrelated. The cutoff grade is the weakest mineralization that can be mined at a profit. Ore reserves are calculated in respect to the amount of ore in place at potential cutoff grades, the tonnages and average grades in identified blocks of ore, and the ultimate tonnage and grade of ore that should be available under projected conditions of recovery and wall rock dilution in mining. The suitability of a deposit for mining is determined in testing and evaluation work related to the physical and chemical nature of the ore, hydrologic conditions, and the needs for ground control. *See* ROCK MECHANICS.

Mine Development

Where high topographic relief allows for an acceptable tonnage of ore above a horizontal entry site, an adit or blind tunnel is driven as a cross-cut to the deposit or as a drift following the deposit from a portal at a favorable location for the surface plant, drainage facilities, and waste disposal. In situations where the deposit lies below or at a great distance from any portal site for an adit, entry must be made from a shaft collar or from an incline or decline portal. A large mine will commonly have a main multipurpose entry and several more shafts or adits to accommodate personnel, supplies, ventilation, communication, and additional production.

Adits. Access by adit generally provides for relatively low-cost underground mining. The broken ore from above the adit level can be brought to the portal in trains, conveyor belts, and rubber-tired trucks without the need for hoisting, and the workings can be drained without pumping. The driving of an adit is generally less expensive per unit distance of advance than the sinking of a shaft or the driving of an inclined access. In areas of low topographic relief and in the mining of deep orebodies, the sinking of a

shaft will often be a more economical approach than the driving and maintaining of a considerably longer incline or adit from a remote part of the site.

Shafts. Production shafts are generally located in stable ground on the footwall side of a dipping deposit rather than in the deposit itself or in the hanging-wall side, where protective pillars would be needed to maintain stability as mining progresses. A shaft may be inclined to follow the dip of the deposit and avoid increasingly longer cross-cuts to the ore at greater depth, but vertical shafts are more common because of their lower construction and maintenance cost per unit of depth and their better efficiency for hoisting ore. Shafts are sunk as rectangular or circular openings 15–30 ft (5–9 m) in diameter; they are equipped with a headframe and hoisting system and are lined with timber, steel forms, or concrete for ground support. Smaller shafts 5–15 ft (1.5–5 m) in diameter, generally for escapeways and ventilation, may be bored by mechanical drilling machines.

Inclines. Inclines equipped with hoists, declines for access by rubber-tired equipment, and gently inclined spiral ramps for diesel-powered truck haulage allow for direct access to relatively deep mine levels without having to transfer the ore and materials to hoisting systems.

Development workings. Development workings in the deposit consist of mine levels and sublevels, with drifts in the ore zone or in the more stable rock on the footwall side of the ore zone. Level workings serve as passageways for miners and low-profile equipment and as haulageways. In broken or unstable ground, passageways and haulageways are supported by timber sets and steel beams or arches; further stabilization is given by rock bolts, sometimes in combination with cable bolting and wire mesh, and the walls may be lined with concrete or spray-on shotcrete.

The raises that connect levels and sublevels provide for the removal of broken ore (chutes and ore passes), for access by miners, and for ventilation and supply routes.

In conventional mining and in the most common development procedures, headings are advanced in a cyclic sequence of drilling, blasting, mucking (removal of broken rock), and installing ground support. In continuous mining, the cycle is replaced by rapid excavation, a single operation in which headings are advanced by powerful tunnel boring and road header machines with teeth that break rock from the face. In situations where the uniformity and texture of the rock and ore permit development by continuous mining, the walls of the resulting passageways are smoother and more stable than would be provided by conventional cyclic operations involving blasting. *See* TUNNEL.

The continuous mining procedure of raise boring is well established. Shaft boring is used in the sinking of small-diameter ventilation shafts and escapeways. The driving of mine level development headings by cutting and boring machinery is more common in coal, potash, and salt deposits and in relatively soft sandstones and shales than in hard ore and rock.

Hydraulic breakers provide successively smaller rock sizes at development headings, and the broken rock is collected at the face by mechanical loading machinery and transferred to the mine haulage system by mobile conveyors or rubber-tired load-haul-dump machines. Haulage beyond the transfer point has been done by electric-powered locomotives with trains of cars but now is increasingly done by rubber-tired electric- or diesel-powered shuttle cars or trucks and by conveyor belt systems. In shaft mines, the broken rock is collected in underground storage pockets and loaded into skips for hoisting to the surface.

The entire sequence in mine development—the advance of headings, breaking of rock, loading, haulage, and hoisting—is increasingly automated. Teleoperated and autonomous machines have become central to every stage in mining, and new mines are developed with the use of geographic information systems (GIS) technology to accommodate the extensive communication systems and mining methods that relate to operations by remote control. *See* GEOGRAPHIC INFORMATION SYSTEMS.

Mining Methods

A fundamental condition in the choice of mining method is the strength of the ore and wall rock. Strong ore and rock permit relatively low-cost methods with naturally supported openings or with a minimum of artificial support. Weaker ore and wall rock necessitate more costly methods requiring widespread temporary or permanent artificial support such as rock bolting. Large deposits with weak ore and weak walls that collapse readily and provide suitably broken material for extraction may be mined by low-cost caving methods. Few mineral deposits are so uniform that a single method can be used without modification in all parts of the mine. Mining to an increasing depth with higher stress conditions and mining from a thicker portion of an orebody into thinner or less uniform portions will especially call for changes in method.

Naturally supported openings. The stopes remain open, essentially by their own strength, during ore extraction. Stability may be maintained to some extent by timbers, rock bolts, and accumulations of broken ore. The workings may collapse with time or may eventually need to be filled with waste material to protect workings in adjacent areas. Backfilling involves the placement of a paste of cemented waste rock or mill tailings. The methods range from gophering, an unsystematic small-scale practice, to carefully planned and executed systems using limits determined by rock mechanics investigations.

Open stoping. This is used in steeply dipping and thin orebodies with relatively strong ore and wall rock. In overhand methods the ore is stoped upward from a sill pillar by miners working on a staging composed of stulls (round timbers) and lagging (planks). With the drilling and blasting of successive small blocks of ore from the back (roof), the broken ore falls onto lower stagings and to the bottom of the stope; it is

collected on the haulage level through draw points or chutes. In underhand stoping the ore is mined downward in a series of benches, and the broken ore is scraped or hauled into a raise or ore pass for collection on a lower mine level. The width of an open stope is limited by the strength of the ore and its capability to stand unsupported. Occasional pillars, generally of waste or low-grade zones in a vein, are left for support; timber stulls may be wedged between the stope walls for stability as well as for access, and rock bolts may also be used to maintain wall stability.

Sublevel stoping. Also referred to as longhole or blasthole stoping, sublevel stoping is practiced in steeply dipping and somewhat wider orebodies with strong ore and strong walls (**Fig. 2**). Sublevel drifts and raises or slots are driven at the ends of a large block of ore so that a series of thinner horizontal slices can be provided. Miners in the sublevels drill patterns of radial holes (ring or fan drilling) or quarrylike parallel holes (slashing). Beginning at the open face of the initial slot, the ore is blasted in successive increments, and the broken ore falls directly to the bottom of the stope. A crown pillar is generally left unmined at the top of the stope to support the next major level.

Vertical crater retreat. This is a method of sublevel stoping in which large-diameter blastholes are drilled in a parallel pattern between major levels, and the ore is broken from the bottom of the stope in a sequence of localized blasts. All of the drilling, loading, and blasting are done by miners and teleoperated machinery in the upper level, so there is no need for access to the ore from below as the stope progresses upward.

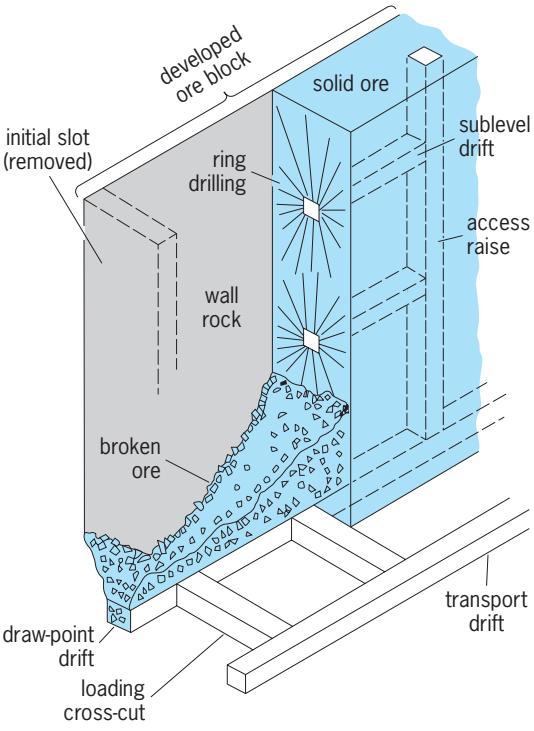


Fig. 2. Sublevel stoping, with ring drilling.

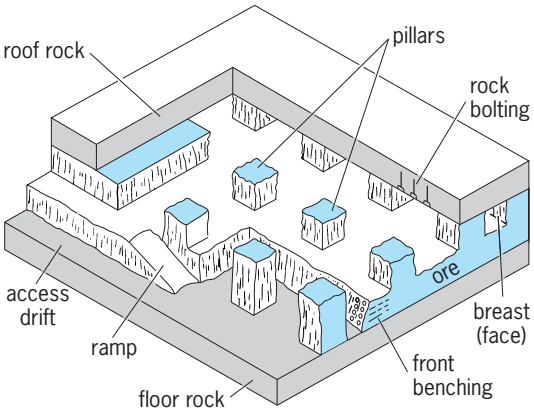


Fig. 3. Room-and-pillar mining; two-stage benching operation.

Room-and-pillar mining. This is also referred to as stope-and-pillar mining when done in a less regular pattern. Room-and-pillar mining is done in coal seams and in flat-lying or gently dipping ore and industrial mineral deposits (**Fig. 3**). It is a low-cost method of underground mining because fast-moving rubber-tired equipment can operate freely, especially in large rooms and haulageways. Thin-bedded deposits are generally mined in a single stage (pass) by conventional or continuous mining; thicker deposits are mined in a two-stage benching operation. In deposits of considerable thickness, an underground quarrying operation follows the first-stage opening of a development level for sufficient access by open-pit-type blasthole drills. Room-and-pillar mining is generally limited to depths on the order of 3000 ft (914 m) in hard-rock mines and to lesser depths in coal mines because of rock bursts and similar manifestations of high-stress concentration on the pillars. Extraction in mining generally amounts to about two-thirds of the ore in a bedded deposit, with the remaining ore being left in pillars; in places where pillars can be "robbed" and the roof allowed to settle, extraction can be increased to 90% or more. *See* ROCK BURST.

Shrinkage stoping. This is an overhand method in which broken ore accumulates in the stope, affording temporary support for the walls and a working platform for miners (**Fig. 4**). Shrinkage stoping is most applicable to steeply dipping veins with strong ore that will stand across a span and with relatively strong wall rock that would slough into the stope in places if left completely unsupported. When ore is broken, it has an expansion or swell factor; this necessitates a periodic drawing (shrinking) of some of the broken ore from the draw points and chutes to allow for continued access to the top of the stope. When all of the ore has been broken except for that left in pillars to protect the adjacent raises and mine levels, the entire content (the magazine) of the stope is drawn. The empty stope may be left open or filled with waste rock, and the pillars may eventually be mined.

Artificially supported openings. In these methods, workings are kept open during mining by using

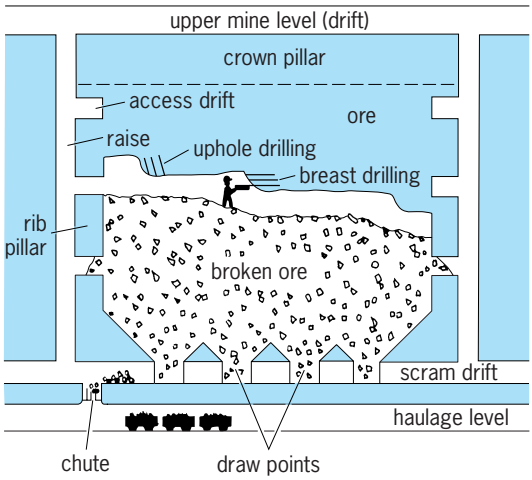


Fig. 4. Shrinkage stoping, longitudinal section.

waste material, timber, and hydraulic props. After the ore is extracted, the workings are filled to maintain stability or are allowed to cave.

Cut-and-fill stoping. This method, also referred to as drift-and-fill, is used in steeply dipping orebodies in which the ore has sufficient strength to be self-supporting but the walls are too weak to stand entirely without support (**Fig. 5**). Most cut-and-fill stoping is done overhand, with the drilling and blasting phase similar to that in shrinkage stoping; the broken ore, however, is removed from each new cut or slice along the back, and the floor of the stope is built up of waste material such as sand or mill tailings brought in by pipeline as a water slurry. The smooth and compacted or cemented fill material provides an especially suitable floor for rubber-tired machinery. Variations in cut-and-fill mining include the ramp-in-stope system, in which load-haul-dump equipment can move rapidly in and out of the stope on an inclined surface of fill material, and the less-mechanized system of resuing in narrow veins. In resuing, ore and waste material are broken separately and the waste material is left to accumulate as fill. One additional system, undercut-and-fill, is applied to bodies of weaker ore. It provides a solid artificial back of reinforced and cemented fill for the mining of successively underlying slices of ore.

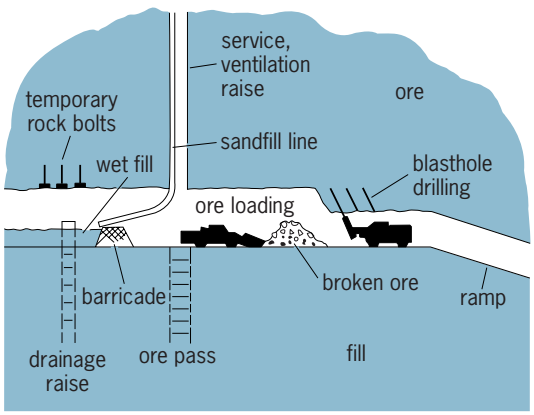


Fig. 5. Cut-and-fill stoping with sand slurry and ramp.

Square set stoping. This is a labor-intensive and high-cost method that has been classically used in situations where the ore is too weak to stand across a wide or long back and the walls are not strong enough to support themselves. A square set, a skeletal box of keyed timbers, is filled and wedged into the available space as each small block of ore is removed by drilling and blasting. Mining continues by overhand or underhand stoping, and the stope becomes a network of interlocked square sets. The sets in the mined portion of the stope are filled with mill tailings or waste rock and pillars are left between mined-out stopes for additional wall support while the remainder of the deposit is being mined. Because of its high cost, square setting is no longer in use; it has been superseded in many mines by cut-and-fill, top slicing, and sublevel caving methods.

Longwall mining. This method is applicable to uniform and extensive but relatively thin deposits. Primarily a highly mechanized and increasingly automated coal mining method at depths where rock pressures are too high for safe room-and-pillar mining, it has also been used in potash deposits and to some extent in bedded iron, copper, and uranium orebodies. In the South African deep gold mines, a form of longwall mining is used in the thin-bedded ore zones.

In longwall mining, practically all of the coal or ore is recovered except for that left in safety pillars to protect surface structures.

The basic practice is to maintain a temporary opening in a uniform line along a working face and then to allow the roof to cave onto the floor or waste fill (gob) behind the active area. In a typical mechanized longwall coal operation, the roof support units are canopies with hydraulic-powered adjustable legs or chocks that are moved ahead as the coal is shaved into slices by shearing and plowing machinery with integrated conveyor systems. In the mining of South African gold reef deposits, longwall-type mining is done by drilling and blasting; the active area is kept open by hydraulic props and timber-concrete packs, and the mined-out areas are filled to some extent by waste rock or cemented mill tailings.

Longwall mining systems allow for a high abutment pressure to build up in solid ore or coal in advance of the face, a low-pressure zone to exist in the working area just behind the face, and a normal lithostatic pressure to build up again in the mined-out and caved or gob-filled area as the face is moved ahead.

Top slice mining. Seldom used today, this method has been applied to wide and steeply dipping deposits with weak ore and weak walls. It has been of use in recovering pillars that have been left between filled stopes. It is a relatively expensive and labor-intensive method with a requirement for abundant timber, but it permits nearly total extraction of the ore. Top slicing is ultimately a caving method of mining, but the ore must first be drilled and blasted, and temporary support is needed between the taking of each successive downward slice or horizontal cut of ore. Working begins in drifts and cross-cuts on a mining floor

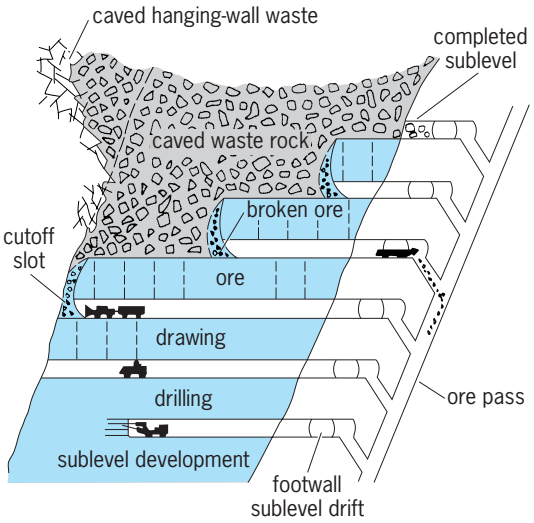


Fig. 6. Sublevel caving, with stages of development and mining.

at the top of a raise; after the driving of a series of adjacent cross-cuts so that a slice of sufficient width has been taken, a mat of timber and scrap lumber is laid down on the floor and the supporting timbers are blasted to cave the overlying rock. A new slice is mined laterally from drifts and cross-cuts under the mat, with the mat supported by timber props (stulls). A mat is again laid down, supports are blasted, and subsequent slices are mined beneath the subsiding accumulation of timber mats and waste rock.

Caving methods. These methods are used in large orebodies with relatively weak ore and with weak wallrock that will collapse as the ore is removed. Geologic conditions must permit subsidence, and the ore must be sufficiently jointed or fractured to form fragments small enough to be handled in drawpoints and raises. Ore recovery in mining is generally quite high, but a certain amount of dilution from waste rock must be accepted.

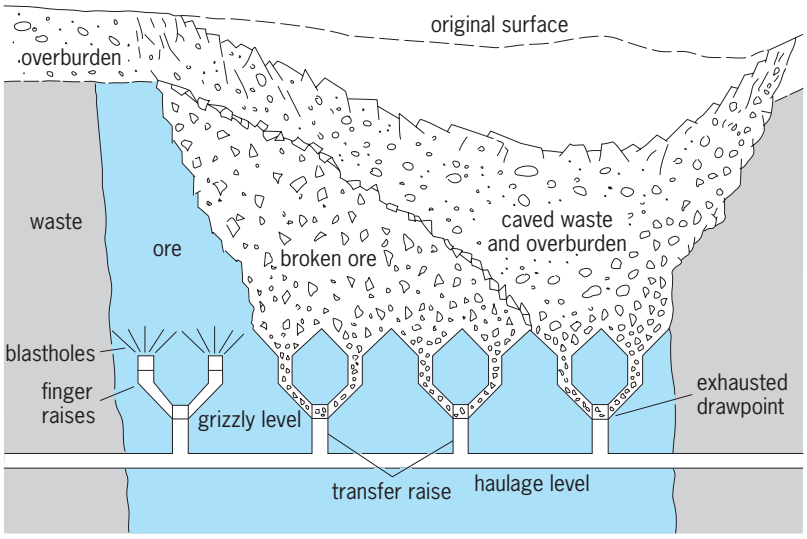


Fig. 7. Block caving, with principal haulage level, driving raises to production (grizzly) levels, and raises to workings.

Sublevel caving. This type is most suited to large and steeply dipping orebodies with weak walls and with ore that has enough stability to maintain sublevels (**Fig. 6**). It is similar to sublevel open stoping, but in this method the walls and the back are allowed to collapse. The ore is mined in downward increments that are drilled, blasted, and drawn from levels below the ore. Access drifts are driven on the footwall side of the orebody, sublevel cross-cuts are driven in ore, and fans of blastholes are drilled at intervals in the cross-cuts. A steplike succession of slices is mined in retreat from the hanging wall, with the wall rock collapsing and following the extraction of the ore. As each fan of holes is blasted, the broken ore caves into the sublevel, where it is loaded and transported to the ore pass. Broken waste rock fills the void as the ore is drawn. When an excess of waste rock begins to dilute the broken ore, the drawing is stopped and the next fan of holes is blasted.

Block caving. This is applied to large and relatively uniform bodies in which both ore and waste will cave readily (**Fig. 7**). Production on the order of 50,000–75,000 tons (45,000–68,000 metric tons) per day can be achieved at a very low mining cost, but the capital cost of a block-caving mine is high. A mine is prepared for block-caving operations by establishing a principal haulage level, driving raises to production levels (slusher or grizzly levels), and driving a larger number of raises to workings on an undercut level beneath the orebody or block to be mined. Caving is initiated by drilling and blasting a slice of ore above the undercut level and, if necessary, by excavating narrow stopes at the boundaries of the block. With the drawing of the initially broken ore, the block begins to cave under its own weight. With further drawing, the entire column of ore and overburden rock continues to subside and break upward for as much as 4000 ft (1220 m) to the surface, where a depression forms. The ore, broken and crushed in caving, flows through cone-shaped draw holes and finger raises. The finger raises are carefully monitored at draw points on the grizzly level so that the caving action is kept uniform and salient channels of subsiding waste rock are not allowed to form prematurely. Broken ore collected from finger raises reaches the haulage level through transfer raises. *See* EXPLOSIVE; MINING; PROSPECTING.

William C. Peters

Bibliography. R. E. Gertsch and R. L. Bullock (eds.), *Techniques in Underground Mining*, SME, 1998; H. L. Hartman and J. M. Mutmansky, *Introductory Mining Engineering*, 2002; H. L. Hartman (ed.), *SME Mining Engineering Handbook*, 2 vol., 1992; B. Stack, *Handbook of Mining and Tunnelling Equipment*, 1982; K. S. Stout, *Mining Methods and Equipment*, 1980.

Underwater demolition

The controlled use of explosives to achieve specific underwater work requiring cutting, fragmenting, perforating, or pounding. The use of underwater explosives was pioneered by the military before and

during World War II for the removal of beach obstructions prior to amphibious landings, mine clearance, channel cutting, and the demolition of wrecked vessels.

With increasing commercial work in the oceans resulting from increased trade, complex marine salvage operations, oil and gas exploitation, and other industrial development, underwater demolition techniques have been widely applied and expanded. The technology has been developed to allow widespread and safe use of explosives as a tool in constructive underwater work.

The type of explosive material, its characteristics, and the shape and size of the charge all determine the effectiveness of explosives as a tool. Explosives are particularly effective under water because the water itself acts to "tamp" the explosion and increase its effect over that of a similar charge on the surface.

Military applications. The most common use of underwater demolition in the military remains the removal of obstacles to amphibious assault. Divers place charges on the obstacles. These charges are designed to cut or break up the obstacles.

When there is too little time or it is too dangerous to move mines or unexploded ordnance, divers place charges on or near the mine. When these charges are set off, the mine explodes and is destroyed.

Military and commercial organizations also use underwater demolition for clearing wrecked ships from harbors or waterways which must be used. In this work, advantage is taken of the ability of properly designed charges to cut or pound. Cutting charges may be used to cut away masts or to cut the ship's plates and internal structure. Pieces of the ship may be lifted out or the remainder may be pounded down with explosives to provide sufficient clearance for ships to pass over.

Another frequent task for underwater demolition that has both military and commercial use is widening or deepening a channel for the passage of shipping. Where channels must be made larger, explosive charges placed in regular patterns may be used to break up hard clay, sand, or rock bottoms so the bottom material can be removed by dredging.

Commercial applications. In addition to the wreck removal and channel widening and deepening uses described above, a wide range of commercial applications has been developed for underwater demolition. In the offshore oil and gas industries, charges may be placed and fired well below the ocean floor to open fissures in the rock, cut off steel pipe, open trenches, and remove old structures.

A large number of new explosives suited to particular tasks and the techniques to employ them have been developed. Special charge shapes and sizes allow very precise work to be done with explosives. While charges for military purposes are usually placed by divers, explosives used commercially can also be placed by crewed submersibles or remote-operated vehicles. *See* DIVING; EXPLOSIVE; SHIP SALVAGE.

William I. Milwee

Bibliography. E. I. du Pont de Nemours and Co., Inc., *Blaster's Handbook*, 1969; R. Larn and R. Whistler, *Commercial Diving Manual*, 1984; *Military Explosives*, 1991; U.S. Navy, Bureau of Ordnance, *Use of Explosives in Underwater Salvage*, 1956.

Underwater navigation

The process of directing the movements of submersible vehicles, and divers, from one point to another. The development of improved submersible vehicles, coupled with advances in saturated diving, has resulted in new requirements for underwater navigation. Various methods which have proved successful include acoustic transponder systems, dead reckoning, surface-referenced navigation from a support ship, terrain map matching, homing, and various combinations of these. The choice of the navigation system depends on such factors as precision required, area to be covered, availability of surface vessels, sea state under which they are expected to operate, and the duration of the mission. Redundant systems are routinely provided to maintain the safety of crewed vehicles. *See* DIVING; SUBMARINE; UNDERWATER VEHICLE.

Most mission requirements must be met by a combination of navigation equipments in an integrated system (**Fig. 1**). Therefore the acoustic equipments described are compatible to a large extent, and interference between sonars is minimized. Further, dead-reckoning computations make provision for periodic updating of position from other navigational equipment.

Acoustic transponders. The most accurate undersea navigation systems use acoustic transponders to precisely determine position with respect to points on the sea floor. A minimum of two transponders are normally dropped from a surface or submerged vehicle to establish a bottom-tethered array. The transponders can be interrogated by surface vessels, submersible vehicles, or divers. When the navigation solution is computed on the surface vessel, the submerged vehicle also carries a transponder, and the position of the submerged vehicle is computed relative to the bottom-tethered array. The location of these transponders can be determined by an accurately positioned surface vehicle that transmits a series of interrogation pulses from different locations. The location of the submersible is determined by a series of round-trip time measurements obtained by interrogating transponders at known locations.

The small size and low power consumption of microminiaturized computers readily allow on-board, and real-time, position determination by a submerged vehicle interrogating the transponder array. This method is widely used with both crewed and crewless submersibles as well as towed vehicles. [Crewless underwater vehicles are customarily referred to as unmanned underwater vehicles (UUVs), and are also referred to as autonomous underwater vehicles (AUVs).] Since the interrogating sonar is well below the surface layers where multipath reflections from the air-water interface do not distort

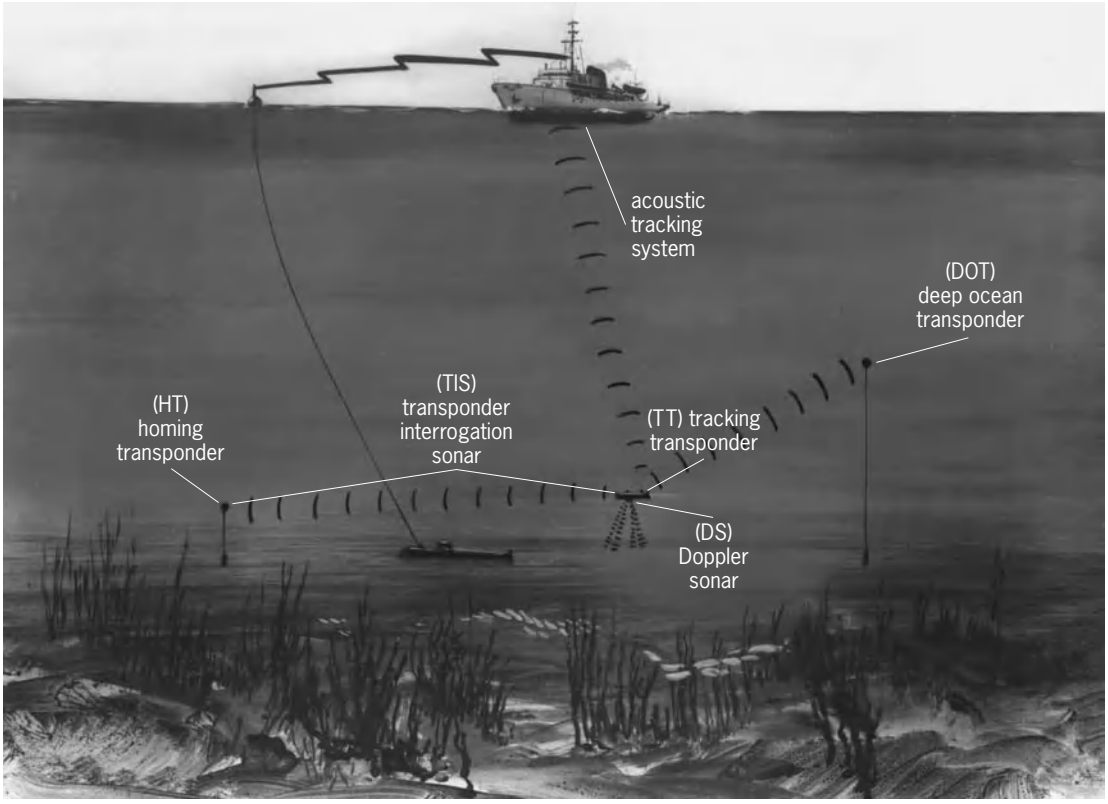


Fig. 1. Acoustic navigation systems, which employ various equipment.

received sonar signals, it provides a high-accuracy solution. The time delay between interrogation and received return (plus a fixed transponder turnaround delay) is a measure of the range of each transponder from the interrogating source. Individual transponders are identified by selected frequencies or digital codes. The differentiation, whether by frequency or code, may be applied to either the interrogate pulse or to the transponder reply. If each transponder has a unique response to interrogation, then all transponders within range can be interrogated with a common signal.

A typical deep-ocean navigation transponder (**Fig. 2**) is of cylindrical configuration and contains the electronics, battery, and transducer. A suitable float and anchor are also provided. Because of the bottom shadow found in deep water (discussed below), it is essential to place the transducer well above the sea floor. The height chosen for the deep-ocean transponder (DOT) is a compromise between range and uncertainty of position due to the watch circle, which is defined as the horizontal range over which a tethered object can vary in response to ocean currents. The frequencies employed for both reception and reply are commonly in the range of 7 to 20 kHz.

The key element of any transponder design is the signal recognition circuit. This circuit must be able to recognize short interrogation pulses of varying amplitude in the presence of widely changing noise levels and to reject a large percentage of false signals due to noise bursts. The most successful circuit consists of a modified differential amplifier and a phase detector to compare the total energy received over a wide bandwidth to that in a relatively narrow band centered on the interrogation frequency. This approach rejects loud noise pulses but responds to weak interrogations. After recognition of a signal, the receiver is gated off for a short period of time to preclude multiple responses, for example, those which might occur from signals reflected off the bottom or the surface of the ocean. *See* DIFFERENTIAL AMPLIFIER; UNDERWATER SOUND.

A triangular array of transponders (**Fig. 3**) is an efficient deployment pattern since the addition of a single transponder can double the area in which the vehicle is able to search or navigate. This doubling is intuitively clear from the following argument. Three transponders define a triangle.

Fig. 2. Deep-ocean navigation transponder. This beacon is about 55 in. (140 cm) in length and 5 in. (12.5 cm) in diameter. (*Sona Tech, Inc.*)

Adding a fourth transponder to form a parallelogram doubles the coverage area. This may be a prime consideration if the submersible must carry and deploy the transponders. The transponder spacing is determined from the estimated maximum range of the transponders. The maximum range is a complex function of vehicle altitude, transponder tether length, and the operating-area depth-temperature profile.

In deep water, acoustic rays are refracted upward by thermal gradients that change the index of refraction for propagation of sound waves, thereby creating a bottom shadow zone. Each two-way time delay measurement from the UUV to a transponder locates the UUV on a sphere that is centered at the transponder location and whose radius is the one-way distance to the transponder. The intersection of two of these spherical loci places the UUV on a circle. The addition of a vertical measurement produces two potential intersections with this circle. This two-point ambiguity limits the range at which a vehicle near the bottom can hear a transponder far more severely than attenuation of the signal by the absorption or spreading loss. Maximum practical range of transponders, such as the deep-ocean transponder is less than 7–10 nautical miles (13–18 km) near the surface and reduces to less than 4 nmi (7 km) when both the transponders and the interrogating vehicle are 300 ft (90 m) above the bottom. The compromise usually chosen is to space transponders 1–3 nmi (2–5.5 km) apart and to place more transponders where the area of operation exceeds twice the transponder spacing.

A two-transponder array is useful if additional data, such as the vertical displacement, are known. Both a three-transponder array and a two-transponder array plus vertical displacement result in a two-point ambiguity. The ambiguity can usually be resolved by a rough knowledge of the vehicle's position.

The transponder array survey-calibration error is the largest contributor to the navigation error. A transponder array is calibrated from either a surface ship or a submersible by making a pass over each transponder (to determine depth) and a pass over the center of each baseline (to determine the intertransponder spacing). Data collected from a submersible have the advantage of being taken below surface layers. The collected data are batch-processed with least-squares fit algorithms to compute the transponder positions. A number of transponders incorporate self-calibrating features. Upon receipt of a calibrate command, the transponders operate in what is commonly referred to as a sing-around mode in which they interrogate one another, and in effect generate (or collect) the data which allow determination of intertransponder spacing. The transponders may additionally be instrumented and programmed to measure pressure (to determine transponder depth) and temperature (to determine sound velocity within the array).

It is possible to solve for position within a transponder net with a hand calculator using a simplified and linearized set of range equations. However, the exact equations are nonlinear and are most expeditiously solved with a digital computer. A computer solution can dispense with linearization and simplifying assumptions, and will readily compute ray-bending corrections and harmonic mean sound velocity, and account for Earth curvature if necessary. A primary advantage of the digital computer is in the implementation of data filters to reject inconsistent transponder returns and navigation solutions. Transponder navigation systems have demonstrated excellent repeatability of data and the ability to return to the same spot on the sea floor to an accuracy of 16 ft (5 m) or less.

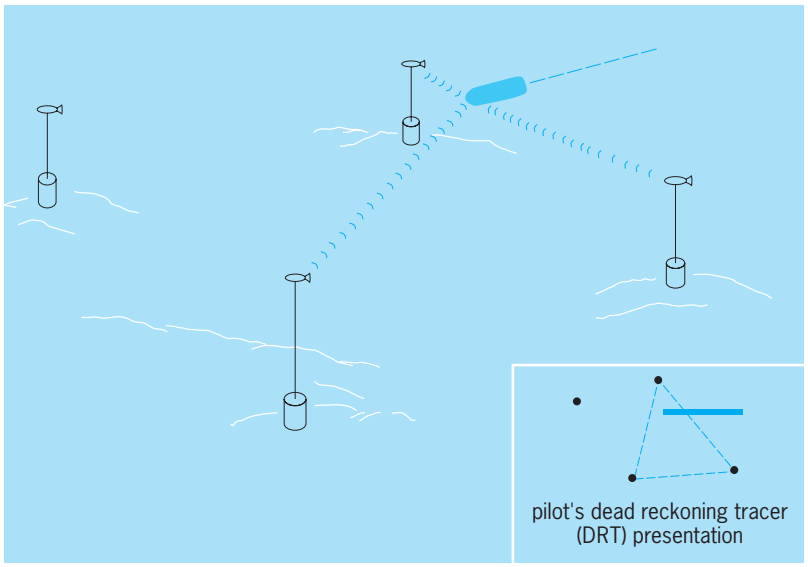


Fig. 3. Submersible navigating from underwater acoustic transponders.

Dead reckoning. An undersea vehicle needs sensors to show distance traveled and direction of travel to mechanize dead-reckoning estimation of position. The most promising sensor for operation near the sea floor is a Doppler sonar. Ground speed, fore-aft and athwartship, can be determined from the Doppler shift in frequency of signals returned from the sea floor. Pulse-type and continuous-wave Doppler sonars are available. The pulse type, which makes use of a frequency tracker circuit to lock on the received pulse, appears to offer the best accuracy, and it has the ability to operate to 600 ft (180 m) above the sea floor. Range gating to reduce noise reception is another feature possible with pulse-type Dopplers. Certain Doppler sonars operate on returns from reverberations from the water mass. This technique extends the usefulness of Doppler sonars to those situations in which the vehicle is out of range of the sea floor. Accuracy is of course much less, since the velocity measurement is referenced to a moving water mass. *See* SONAR.

Distance traveled is calculated from integration of the x and y sonar velocities. Accuracies of 0.1–0.2% of distance traveled can be obtained at altitudes ranging from about 3 to 600 ft (1 to 180 m) above the bottom. Much greater range is obtained in deep-water Doppler sonars. The technology of Doppler velocity

logs (DVLs) has been significantly improved since the mid 1990s, in part because of the increased interest in and military utility of UUVs.

Correlation velocity sonar can provide good velocity data at much greater bottom ranges. This sonar transmits pulsed energy in a broad beamwidth vertically to the bottom. Reflected echoes are received by an array of identical hydrophones, each having a broad beamwidth. The C-shaped receiving array provides fore-aft and athwartships components of velocity which are estimated independently of the velocity of sound. A correlation sonar, operating at altitudes of up to 16,000 ft (5000 m) over the sea bottom, can measure velocity components to accuracies which, when integrated, give a positional accuracy of 0.4% of distance traveled.

Since the Doppler and correlation sonars both measure in ship's body axes, a north-referenced heading measurement is required for navigation. This is usually provided by a small gyrocompass or magnetic compass. Accuracy for the dead-reckoning system is about 0.5% of distance traveled. *See* GYROCOMPASS.

In addition to Doppler sonar, some specialized submersibles carry inertial dead reckoning in the form of a small, stabilized platform. The accuracy of any dead reckoning system can be improved by receiving additional position information that is more accurate than the errors that have accumulated since the last position update. This opens up opportunities to use sonar-based underwater terrain map matching or occasional Global Positioning Satellite System (GPS) fixes collected by floating an antenna to the surface as sources of these independent position updates. *See* DEAD RECKONING; INERTIAL GUIDANCE SYSTEM.

Homing and piloting sensors. Most submersible vehicles carry additional navigation sensors. The horizontal obstacle sonar (HOS) which is a constant-transmission frequency-modulated sonar (CTFM), is normally used to detect objects ahead of the submersible (**Fig. 4**). This sonar also has a transponder channel for determining bearing and range to specially designed transponders. An altitude-depth sonar provides vertical navigation by furnishing depth and altitude off the bottom (Fig. 4). Finally, a vertical obstacle sonar (VOS) determines heights of objects in the path of the vehicle (Fig. 4). This sonar is a constant-transmission frequency-modulated type with a higher transmission frequency than the horizontal obstacle sonar. Both horizontal and vertical obstacle-avoidance sonars are useful in under-ice navigation. *See* HOMING; PILOTING.

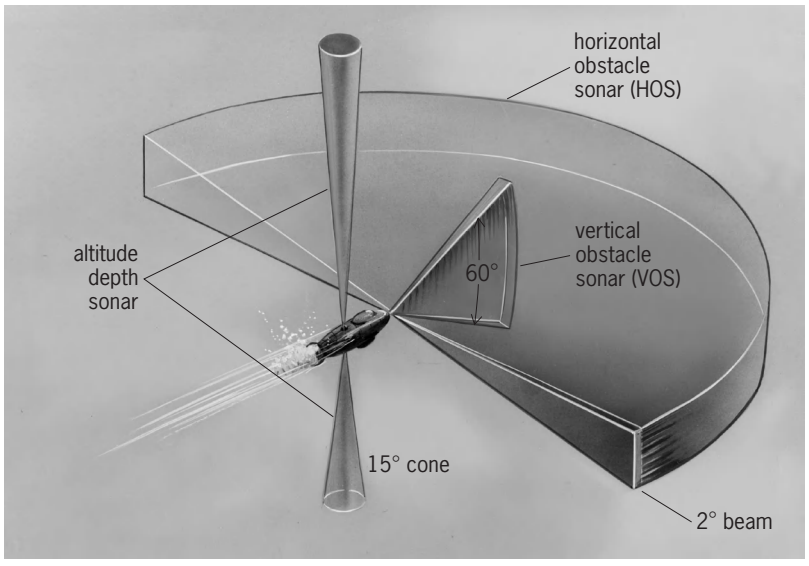


Fig. 4. Configurations of three types of pilotage sonars.

Surface reference. An acoustic tracking system allows the monitoring and vectoring of the position of a submersible from a surface vessel, where space and weight are not at a premium. The Global Positioning Satellite System or other radio techniques can be used to determine the ship's own position in geographic coordinates. Surface tracking systems are of two types: ultrashort baseline, employing orthogonal line hydrophone arrays; and short baseline, employing four separate hydrophones. *See* MARINE NAVIGATION.

When monitoring a vehicle carrying a transponder (Fig. 1), both range and direction (bearing and depression angle) can be determined. Range is computed from the time interval between interrogation and received reply, while bearing and depression angles are determined from phase differences between the received signals at different elements of the hydrophone array (first system), or from time-of-arrival differences at the separate hydrophones (second system).

The ultrashort-baseline system employs a single transducer, mounting two orthogonal arrays in an assembly less than 10 in. (25 cm) in diameter. Typical systems have achieved accuracies of 1% of slant range, within a horizontal offset of four times the water depth.

The short-baseline system consists of an array of four hydrophones mounted on a ship in two orthogonal baselines. The length of each baseline is the longest possible consistent with the ship's geometry. The hydrophones are mounted in wells and frequently can be extended below the ship's hull to reduce multipaths and ship's noise. Similar mounting is employed for the single transducer of the ultrashort-baseline system.

Since it employs time-of-arrival measurements rather than phase comparisons, the short-baseline system is less accurate than the ultrashort-baseline system, but it is usable at greater depths and over wider areas. Sometimes both systems are combined in a single installation.

The surface ship has one significant disadvantage: its acoustic equipment is located in the surface layers and is most subject to ship's noise and nonhomogeneities in the propagation medium. Therefore surface-ship tracking usually augments rather than replaces those systems carried by submersible vehicles.

Submarine navigation. Submarines must operate over wide areas of the ocean and often under highly secure and covert conditions; therefore, navigation

techniques which depend upon sonic or electromagnetic emissions, or which restrict movements of the vehicle to either near-bottom or near-surface regions, are judged to be too restrictive. The ability of inertial navigation systems to operate without frequent recourse to external position updates makes them prime candidates for submarine navigation. The most stringent navigation requirements for navigation accuracy are associated with missile-firing submarines. Contemporary sea-launched ballistic missiles (SLBMs) navigate using self-contained sensors (such as inertial sensors and star trackers). Inasmuch as they mechanize a dead-reckoning navigation solution their impact errors are sensitive to position and velocity errors at the time of launch. Improvements of other aspects of missile guidance systems have driven requirements to improve the ship's navigation system. The additional need for covert and secure long-duration navigation has resulted in the evolution of a navigation configuration that uses a pair of electrostatic gyroscope navigators (ESGNs) as the primary elements of an integrated navigation subsystem. The very stable, low-drift characteristics of the electrostatic gyroscope result in an inertial navigation system with a low and highly predictable error growth. Weapons system accuracy is further enhanced by velocity derived from both gyroscopes and secure correlation velocity sonar techniques to give a direct measure of the submarine ground speed. Vertical deflection maps, which are used for vertical-axis tilt compensation of the gyroscopes, are generated from combined satellite and oceanographic surveys of the Earth's gravity field. The relatively rare position resets are selected from either Global Positioning System satellite or bathymetric sonar data. The passive electromagnetic-log, which measures vehicle velocity relative to the water and is therefore subject to ocean current errors, is used primarily to damp the gyroscope computational (Schuler) oscillations. The various data are processed in a central navigation computer which uses statistical estimation algorithms. *See* ESTIMATION THEORY.

A class of inertial systems that was originally developed for commercial aircraft has found application in undersea vehicle navigation. The inertial components in these systems, both gyroscopes and accelerometers, are fixed to the vehicle frame in a strapped-down configuration. The mechanical gimbals and torque motors, which had traditionally isolated both gyroscopes and accelerometers from vehicle base motions, have been replaced with mathematical transformations implemented in a digital computer. Additionally, mechanical spinning-wheel gyroscopes have been replaced with the counter-rotating light beams of ring-laser gyroscopes. These solid-state configurations make extensive use of microprocessor and microminiature technology. The low power consumption, small volume, light weight, and low cost of ring-laser-gyroscope strapped-down systems makes them attractive for crewless undersea

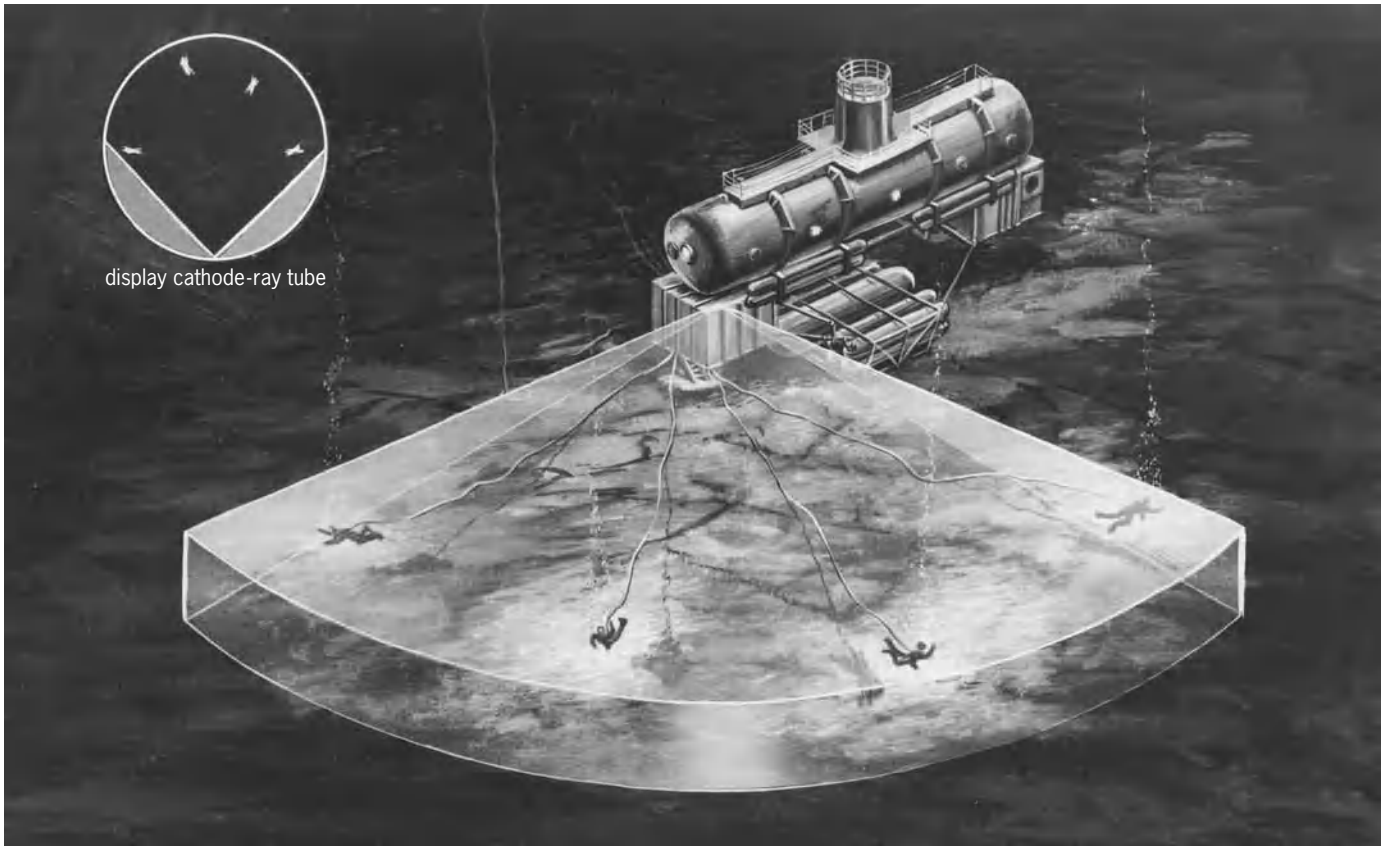


Fig. 5. Sonar for tracking of divers from an underwater habitat. Display cathode-ray tube shows positions of divers to diving officer.

vehicles, which have limited on-board battery-supplied power. Such vehicles do not have the long-duration requirements of nuclear submarine missions, and can therefore use less precise inertial navigators with external aids such as Doppler or correlation velocity sonar. *See* ACCELEROMETER; GYROSCOPE.

The increase in speed, capability, and capacity of microprocessors and computer memory, combined with corresponding decreases in size, power consumption, and cost has permitted UUVs to be given capabilities ordinarily reserved for larger crewed submarines. System reliability is approaching absolute, with increased use of fault-tolerant processor configurations and redundant sensors. The more sophisticated navigation system configurations make use of such computationally intensive techniques as bottom terrain matching, visual image processing, and object recognition. *See* FAULT-TOLERANT SYSTEMS; IMAGE PROCESSING; MICROPROCESSOR; SEMICONDUCTOR MEMORIES.

Diver navigation and tracking. Navigation has always been a difficult problem for divers. With the advent of saturation diving, undersea habitats, and transfer capsules to allow divers to work anywhere on the continental shelves, precise navigation is even more essential. A diver cannot surface from great depths without many hours or days of decompression. Therefore, before humans can be turned loose without restraining lines, there must be positive methods of diver navigation and tracking.

Some of the same sonars developed for submersible pilotage offer an answer for a diver. A horizontally swept constant-transmission frequency-modulated sonar similar to that previously described (**Fig. 5**) is well matched to the excursions of saturated divers, who are limited in allowable vertical movement. By communication with the habitat, divers who stray too far or lose their orientation can be guided back to safety.

Transponder systems similar to those described above have been employed to guide divers and uncrewed submersibles to and from work sites. The transponders, intended for operation at short ranges (up to 1000 ft or 300 m), transmit and receive in the range 100–400 kHz at low power levels. They are consequently much smaller and lighter in weight than typical deep-ocean navigation transponders (Fig. 2).

Charles J. Hrbek; Joseph A. Cestone; Emery St. George, Jr.; Richard L. Greenspan

Bibliography. W. E. Bradley, Evolution of integrated navigation for deep ocean surveying, *Oceans '90,* pp. 451–457, IEEE, 1990; J. A. Cestone et al., Latest highlights in acoustic underwater navigation, *Navigation,* 24(1):7–39, Spring 1977; D. E. DiMasa and W. K. Stewart Jr., Terrain-relative navigation for autonomous underwater vehicles, *Oceans '97*, pp. 541–546, IEEE, 1997; R. L. Greenspan, Inertial navigation technology from 1970–1995, *Navigation,* 42(1): 165–185, 1995; S. K. Hole, B. Woodward, and W. Forsythe, Design constraints and error analysis of the temporal correlation, *IEEE J. Oceanic Eng.*, 17(3): 269–279, 1992; E. Levinson and R. Majure, Accuracy enhancement techniques applied to the marine ring laser inertial navigator (MARLIN), *Navigation,* 34(1):64–86, Spring 1987; J. S. Stambaugh and R. B. Thibault, Navigation requirements for autonomous underwater vehicles, *Navigation,* 39(1):79–92, Spring 1992; J. A. Strickrott and S. Negahdaripour, On the development of an active vision system for 3-D scene reconstruction and analysis from underwater images, *Oceans '97*, pp. 626–633, IEEE, 1997; M. Uliana, F. Andreucci, and B. Papalia, The navigation system of an autonomous underwater vehicle for Antarctic exploration, *Oceans '97*, pp. 403-408, IEEE, 1997.

Underwater photography

The techniques involved in using photographic equipment underwater. Underwater photography is used to document subjects in fresh-water lakes and quarries, in temperate seas, beneath the polar ice pack, and in the deep sea, but by far the greatest percentage of underwater photography is done within sport-diving limits in the tropical oceans. The color and variety of marine life in the Caribbean Sea, Red Sea, or Indo-Pacific Ocean are diverse and compelling, creating an ever-changing portfolio of subjects to photograph.

Scientists use underwater photography to document issues, such as the health of the coral reefs, while photojournalists use underwater photography as a medium to report on the attributes of dive sites. However, the largest growth in underwater photography has come from recreational scuba divers.

Underwater photographers are faced with specific technical challenges. Water is 600 times denser than air and is predominantly blue in color. Depth affects light and creates physiological considerations for the photographer. As a result, underwater photography requires an understanding of certain principles of light beneath the sea.

Principles. As in all photography, consideration of the variables of light transmission is crucial to underwater photography. When sunlight strikes the surface of the sea, its quality and quantity change in several ways. *See* LIGHT; PHOTOGRAPHY.

Refraction. As light travels from air to a denser medium, such as water, the light rays are bent (refracted); one result is magnification of underwater objects by one-third as compared to viewing them in air. The magnification effect must be considered when estimating distances underwater, which is critical for both focus and exposure.

When using a camera equipped with a flat port (outer lens), the refractive index of the lens matches the perspective of the photographer when viewing the scene through the flat port of a facemask. Therefore, a subject that looks to be 3 ft (0.9 m) away to the photographer is actually 3 apparent feet from the lens. Apparent distances may be imprecise, and depend on a subjective guess at how distant a subject may be, but few subjects tend to hold still long enough for the distance to be measured. In this

example of 3 apparent feet, the actual measured distance would be 4 ft (1.2 m). Given the spontaneous nature of underwater photography, the popular convention has always been to work in apparent distances, and lenses are made specifically for underwater use and calibrated in apparent feet. This makes accurate distance estimation a valuable underwater photographic skill.

Using a dome port (correction lens) on the camera nullifies the effects of refraction and magnification and restores the true underwater distance. A dome port also creates a "virtual image," which typically exists a distance of about twice the diameter of the dome from the film plane. An 8-in. (20-cm) dome port will have a virtual image about 16 in. (40 cm) away, and the lens must be able to focus on this virtual image. With some lenses, a diopter (closeup lens) must be added to assure close-focus capability, for with a dome port, if the lens cannot focus on the virtual image, it will not be able to focus at all. *See* REFRACTION OF WAVES.

Absorption. Light is absorbed when it propagates through water. Variables affecting the level of light penetration include the time of day (affects the angle at which the sunlight strikes the surface of the water); cloud cover; clarity of the water; depth (light is increasingly absorbed with increasing depth); and surface conditions (if the sea is choppy, more light will be reflected off the surface and less light transmitted to the underwater scene).

Selective color filtration. Depth affects not only the quantity of light but also the quality of light. Once light passes from air to water, different wavelengths of its spectrum are absorbed as a function of the color of the water and depth. Even in the clearest tropical sea, water serves as a powerful cyan (blue-green) filter. Natural full-spectrum photographs can be taken only with available light in very shallow depths. In ideal daylight conditions and clear ocean water, photographic film fails to record red at about 15 ft (4.5 m) in depth. Orange disappears at 30 ft (9 m), yellow at 60 ft (18 m), green at 80 ft (24 m), and at greater depth only blue and black are recorded on film. To restore color, underwater photographers must use artificial light, typically from a submersible strobe (**Fig. 1**). *See* SEAWATER.

Water also absorbs and changes the quality of the light discharged by the strobe. Light is absorbed not only vertically (as a function of depth) but also horizontally (as a function of strobe-to-subject distance). For effective strobe photography, the photographer should be no more than 10 ft (3 m) from the subject, and ideally within 4 ft (1.2 m). Getting closer maximizes the color and resolution of the underwater photo, and thereby minimizes the number of lenses needed for underwater work. Telephoto lenses, for example, perform poorly underwater because their minimum focus distance puts too much water between the photographer and subject. *See* STROBOSCOPIC PHOTOGRAPHY.

Camera systems. The water column between photographer and subject degrades both the resolution of the image and the transmission of artificial light (necessary to restore color). Therefore, the most effective underwater photos are taken as close as possible to the subject, thereby creating the need for a variety of optical tools to capture subjects of various sizes within this narrow distance limitation. An above-water (topside) photographer might choose to shoot a full-figure photo of a person by using a 50-mm lens from 12 ft (3.6 m) away or by using a 24-mm wide-angle lens from 5 ft (1.5 m) away; the underwater photographer has fewer options. A subject 12 ft away will not reflect enough strobe light for an effective photo, so the photographer must use a wide-angle lens and work closer (**Fig. 2**). In this case, perhaps the best choice would be a 20-mm lens from 3 ft (0.9 m) away.

Since camera lenses cannot be interchanged underwater without flooding the camera, the photographer generally restricts the photographs to those appropriate to the lens. Alternatively, a photographer may carry several camera systems underwater, each dedicated to a specific focal-length lens. Increasingly, housing manufacturers are creating port options for using zoom lenses underwater. Two popular focal lengths are the 17- to 35-mm range (for wide-angle and reef scenics) and the 70- to 180-mm macro range (for fish portraits). Primary lenses still have a wider acceptance among photographers, but the quality of the zooms and the ergonomics of the new

Fig. 1. Dolphin and underwater photographer—13-mm lens. The extreme wide angle of the 13-mm lens (170° diagonal coverage) was used here to capture a shot of an Atlantic bottlenose dolphin and diver. By using an extreme wide angle, it is possible to cover large subjects from a close distance, thereby maximizing color and resolution. (*Photograph by Stephen Frink*)

Fig. 2. Over/under shot, Solomon Islands—16-mm lens. The over/under shot (one-half above the water and one-half below) is best done with a housed camera and dome port. An extreme wide-angle lens is typically used, although sometimes a split diopter is also used to account for the difference in refractive index and exposure values between the air and water. (*Photograph by Stephen Frink*)

generation of housings and ports make variable-focal-length zoom lenses a viable option.

There are two types of underwater cameras—amphibious and housed. Amphibious cameras may be used either underwater or topside, although some lenses are for underwater use only (known as water contact lenses) [**Fig. 3**]. A housed camera is a conventional above-water camera that has been protected from the damaging effects of seawater by a waterproof enclosure. The amphibious camera is protected by a series of O-rings, primarily located at the lens mount, film loading door, shutter release, and other places where controls are necessary. The O-rings make the system not only resistant to leaks but also impervious to dust or inclement weather when used above water.

The housed system is typically larger and bulkier than the amphibious system, but the trend in modern housings is to create a small, compact housing with optimal ergonomic placement of controls. A much wider range of lenses may be accommodated in underwater-housed systems, and custom ports are available specifically for the lenses. O-rings are also used in housings to prevent water intrusion. Some housings have mechanical levers that operate the important camera controls such as shutter release, film advance (although most housed cameras are equipped with motor drives), aperture, focus, and zoom control, although some housings now incorporate electronic controls for some functions. Clearly, the most popular camera for use with underwater housings is the single-lens reflex 35-mm, equipped

Fig. 3. Underwater photographer on a shipwreck—15-mm lens. Shipwrecks provide excellent subjects for wide-angle photography, and including a diver as an element of composition can add color to an otherwise monochromatic subject, as well as a sense of scale. (*Photograph by Stephen Frink*)

Fig. 4. Blue-cheeked butterflyfish, Red Sea—28-mm lens. Colorful tropical marine life is a fascinating subject for the underwater photographer. The challenge is to capture the composition in a natural, nonintrusive fashion. (*Photograph by Stephen Frink*)

Fig. 5. Lionfish, Red Sea—60-mm macro lens. When capturing closeup photographs of subjects that might be frightened by using framers (closeup kit or extension tubes), the housed single-lens reflex camera is the perfect tool. Not only can the photographer achieve precise composition by viewing through the lens, but also modern autofocus technology makes it easy to get a sharp picture. (*Photograph by Stephen Frink*)

with either a macro focusing lens (either a 60-mm or 105-mm macro) or a wide-angle lens (**Fig. 4**). Autofocus is one of the new technologies available in topside cameras that have contributed greatly to the ease of underwater photography.

Lenses. The standard lens for the amphibious camera is either a 35-mm or 28-mm, and at the optimal 3-ft (0.9-m) focus distance these lenses are typically used to photograph coral-reef marine life or head-and-shoulder views of divers. Conventional telephoto lenses are of little use underwater, but macro telephoto lenses (say, 105-mm or 200-mm) permit the photographer to capture small and typically skittish subjects from a greater distance, thereby not alarming them (**Fig. 5**).

For closeup documentation of small subjects, most amphibious systems provide an auxiliary diopter (supplementary lens) and focusing wand that works in conjunction with the normal lens. The diopter is placed in front of the lens, and the focus wand plus framer (an open frame that attaches to and extends from the lens, within which the subject is kept to ensure proper focus) shows what the lens sees. The advantage to this type of system is that it can be added and removed while underwater, so that if a larger fish swims by, the photographer has the option to remove the closeup kit and shoot. Another option (for typically greater magnification) is the use of an extension tube. This machined aluminum tube is inserted between the lens and the camera body (O-ring sealed), thereby changing the focal length of the lens to a magnification ratio of one-third life size (1:3) to twice life size (2:1), depending on which combination of tube and framer is used. A housed camera may use continually focusing macro lenses (closeup) to capture subjects with the standard range, but also with magnification ratios of 1:1.

In order to photograph large underwater subjects and panoramas, a wide-angle lens is necessary (**Fig. 6**). For both amphibious and housed systems, that means an angle of coverage ranging from 94° (equivalent to a 20-mm lens topside) to 180° (equivalent to a 16-mm full frame fisheye lens). The advantage to these optics is that the photographer can be within 3 ft (0.9 m) of a massive subject, yet still gain the color from the artificial light and enhance the resolution by minimizing the water column.

Fig. 6. Caribbean reef shark and diver—16-mm lens. The extraordinary depth of field of the full-frame fisheye lens is apparent here, as both the shark and the diver are in sharp focus. A little perspective distortion is evident: because the shark is nearer to the lens relative to the diver, it seems even larger and more impressive. (*Photograph by Stephen Frink*)

Modern amphibious cameras are constructed of materials that resist corrosion in salt water, such as anodized aluminum or marine-grade plastics for the body, while the best optics are made of multicoated glass designed for water contact. The interface between the glass and seawater are part of the optical formula with many amphibious lenses, particularly the wide-angle lenses, and they do not work well above water as a result. Housings are typically constructed of either anodized aluminum or polycarbonate, both of which are strong and corrosion-resistant.

One advantage of the housed system is that an existing camera system may be modified for underwater use, as opposed to buying a camera and lens system specifically dedicated to underwater photography. Since most cameras used in housings are single-lens reflex and autofocus, the level of sophistication in housed systems is quite high. Some of the existing viewfinder-type amphibious cameras are a bit more simplistic, and require the photographer to compose through a relatively imprecise viewfinder. Housings typically have some means of enlarging the single-lens reflex view for viewing through a facemask.

The primary advantages of the amphibious camera are its compact size (as compared with the housing) and ease of operation. The range of accessory lenses available for amphibious cameras tends to be extensive, covering most underwater imaging opportunities. Once familiar with the amphibious system, the photographer can usually react quickly and accurately to capture the underwater scene. There are certain small subjects that may be too skittish to capture with framers, but other than that, most of the subjects that a photographer could capture with a housing could also be captured with the amphibious camera. There are trade-offs between the ease of operation and small size of the amphibious system, as contrasted with the compositional accuracy of the single-lens reflex in a housing; yet successful photographs are commonly taken with both systems.

Stephen Frink

Deep-sea underwater photography. Photography at depths beyond the limits of sport diving—approximately 150 ft (35 m)—requires the design and use of special camera and lighting equipment. Watertight cases are required for both camera and light source, and they must be able to withstand the pressure generated by the sea. For each 33 ft (10 m) of depth, approximately one additional atmosphere ($\sim 10^2$ kilopascals) of pressure is exerted. At the greatest ocean depths, about 40,000 ft (12,000 m), a case must be able to withstand 17,600 lb/in.^{2} (1200 kg/cm^2). The windows for the lens and electrical seals must also be designed for such pressure to prevent water intrusion. Examples of such deep-sea work are the bathypteroid fish photographed at 4000 ft (1200 m) in the Gulf of Mexico (**Fig. 7**) and a view of the Romanche Trench at a depth of 25,000 ft (7625 m; **Fig. 8**).

Cylindrical cases are most common since they are stronger than rectangular ones. The cameras are typically adapted to accept the standard 100-ft (30-m) motion picture spool of 35-mm film that allows 800 individual 35-mm frames to be exposed. Thin-base cronar film permits even more exposures with cameras that can accept 400 ft (122 m) of this film type. This length allows up to 3200 photographs to be taken before changing film, an important consideration since much time is consumed in raising and lowering the camera gear from the sea.

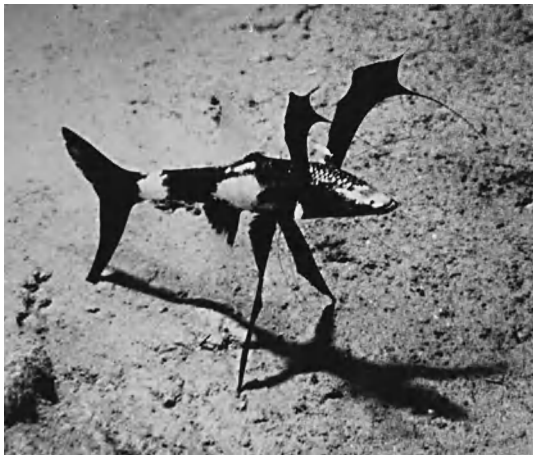


Fig. 7. Bathypteroid fish photographed from the Westinghouse diving vehicle *Deepstar* at 4000 ft (1200 m) depth in the Gulf of Mexico. Tungsten lighting was used. (*Photograph by R. Church*)

Auxiliary lighting is required, since daylight is absorbed in both intensity and hue. The electronic flash utilizing a xenon tube is almost universally used for deep-sea photographs since it is efficient and compatible in color temperature with daylight film. Corrected optics for the camera may be used behind the lens port to compensate for the different refractive index of air and water.

The camera must be positioned and triggered to render the desired photograph, and the great depths preclude a free-swimming human operator. Operation is often from a cable via sonar sensing equipment

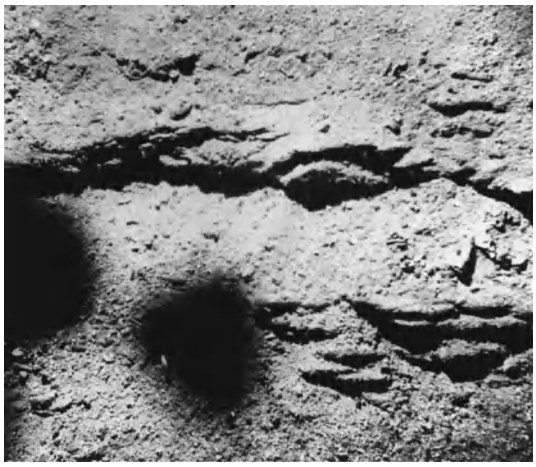


Fig. 8. Photograph taken (August 1956) in Romanche Trench, 0°10′S, 18°21′W, at depth of 25,000 ft (7625 m). A nylon cable was used from the deck of the *Calypso*. (*National Geographic Magazine, March 1958*)

or from deep-diving underwater vehicles. Bottom-sensing switches can operate deep-sea cameras for photographing the sea floor, and remotely operated vehicles (ROVs) can incorporate both video and still cameras. The ROV is tethered to either the surface or the submersible vehicle by an umbilical cable, and it is moved by its own propellers. The operator can direct the movement of the ROV, view the scene through the video camera, and record appropriate still images when the ROV is properly positioned. *See* SONAR; UNDERWATER VEHICLES.

Many deep-sea cameras have an internal data chamber that is photographed simultaneously with the deep-sea exposure. Generally, photographic digital data of 16 digits can record such information as the second, minute, and hour that the photograph was taken, as well as the compass orientation of the photograph and camera-to-subject distance. The camera-to-subject distance is fed into the camera from a sonar sensor mounted on the camera. When the camera-to-subject distance and focal length of the lens are known, it is possible to calculate the field of view for each photograph. Thus, the size of the objects in the photograph may be measured. When the data and the image on the film are inseparable because they are rendered with a simultaneous exposure, the chances of error are greatly reduced.

When an observer descends to great depths in a diving vehicle, the camera can assist in documentation by recording what is seen. Furthermore, the visual data will assist in accurate description of the observed phenomena. Elapsed-time photography with a motion picture camera in the sea is important in studying sedimentation deposits caused by tides, currents, and storms. Similarly, the observation of biological activity taken with the elapsed-time camera and then speeded up for viewing may reveal processes that cannot ordinarily be observed. *See* DIVING; PHOTOGRAPHY; UNDERWATER TELEVISION.

Harold E. Edgerton; Stephen Frink

Bibliography. J. Church, *Jim Church's Essential Guide to Composition*, 1998; J. Church, *Jim Church's Essential Guide to Nikonos Systems*, 1998; M. Edge and I. Turner (eds.), *The Underwater Photographer*, 1999; S. Frink, *Wonders of the Reef: Diving with a Camera*, 1996; A. Kohler and D. Kohler, *The Underwater Photography Handbook*, 1999; J. Liburdi and C. Sherman, *The New Guide to Sea & Sea*, 1998; M. Webster and G. Bradford (illustrator), *The Art & Technique of Underwater Photography*, 1999; N. Wu, *How To Photograph Underwater* (*How To Photograph*, vol. 2), 1994.

Underwater sound

The production, propagation, reflection, scattering, and reception of sound in seawater. The sea covers approximately 75% of the Earth's surface. In terms of exploration, visible observation of the sea is limited due to the high attenuation of light, and radar has very poor penetrability into salt water. Because of the extraordinary properties that sound has in the sea, and because of some of the inherent characteristics of the sea, acoustics is the principal means by which the sea has been explored. *See* OCEAN.

Absorption. Sound has a remarkably low loss of energy in seawater, and it is that property above all others that allows it to be used in research and application. Absorption is the loss of energy due to internal causes, such as viscosity. Over the frequency range from about 100 Hz (cycles per second) to 100 kHz, absorption is dominated by the reactions of two molecules, magnesium sulfate ($MgSO_4$) and boric acid [$B(OH)_3$]. These molecules are normally in equilibrium with their ionic constituents. The pressure variation caused by an acoustic wave changes the ionic balance and, during the passage of the pressure-varying acoustic field, it cannot return to the same equilibrium, and energy is given up. This is called chemical relaxation. At about 65 kHz magnesium sulfate dominates absorption, and boric acid is important near 1 kHz. Absorption has been measured in the laboratory and at sea (**Fig. 1**). The absorption coefficient α is the exponential loss in intensity of an acoustic signal of a given frequency per meter of its path. The ratio of intensities after traversing a distance r is given by Eq. (1). Often absorption

$$I/I_0 = e^{-\alpha r} \tag{1}$$

is defined as the power to the base 10, that is, $I/I_0 = 10^{-\alpha r}$. Due to the very low absorption coefficients

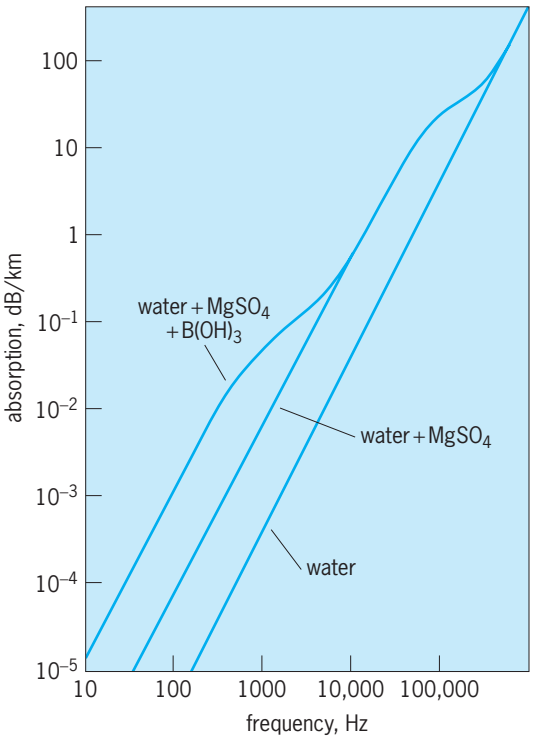


Fig. 1. Sound absorption coefficient as a function of frequency for seawater. For this case, the salinity is 35‰, temperature 4°C (39°F), pressure 100 kPa (approximately atmospheric pressure), and pH 8. Note the effect of the presence of $MgSO_4$ and $B(OH)_3$. (*After F. H. Fisher and V. P. Simmons, Sound absorption in sea water, J. Acous. Soc. Amer., 62:558–564, 1977*)

in seawater, the convention is to change its units to km^{-1}. *See* SEAWATER; SOUND; SOUND ABSORPTION.

Both magnesium sulfate and boric acid are present in very small amounts in seawater. If their concentrations were a few orders of magnitude larger, sound measurement would be as ineffective as it is with radar or optics.

Sound speed. The speed of sound in seawater and its dependence on the parameters of the sea, such as temperature, salinity, and density, have an enormous effect on acoustics in the sea. There have been several empirical equations published in the literature, differing by small amounts, but for the purposes here, the differences are unimportant. To the lowest order in temperature, salinity, and pressure is Eq. (2),

$$C = 1449.2 + 4.6T + 1.34(S - 35) + 0.016Z \quad (2)$$

where the speed of sound in seawater (C) is in meters per second, temperature (T) is in degrees Celsius, salinity (S) is in parts per thousand (‰), and depth (Z) is in meters.

Ocean environment. Generally the environmental parameter that dominates acoustic processes in oceans is the temperature, because it varies both spatially and temporally. The dynamics of the sea are very complex, and one driven by solar heating, winds, atmospheric forces, bathymetry, major ocean currents, turbulence, and the Earth's rotation. It is impossible to include all of these effects in the equations of acoustics. But by limiting the theory to the most salient features, and through reasonable approximations, some solutions can be obtained that are in agreement with experimental results. When agreement is not satisfactory, either the acoustic theory or the ocean model must be modified.

Solar heating of the upper ocean has one of the most important effects on sound propagation. As the temperature of the upper ocean increases, so does the sound speed. Winds mix the upper layer, giving rise to a layer of water of approximately constant temperature, below which is a region called the thermocline. Below that, most seawater reaches a constant temperature. All these layers depend on the season and the geographical location, and there is considerable local variation, depending on winds, cloud cover, atmospheric stability, and so on. Shallow water is even more variable due to tides, fresh-water mixing, and interactions with the sea floor. Major ocean currents, such as the Gulf Stream and Kuroshio, have major effects on acoustics. The cold and warm eddies that are spun off from these currents are present in abundance and significantly affect acoustic propagation. *See* GULF STREAM; KUROSHIO; OCEANOGRAPHY.

Units. The science of underwater sound is the study of pressure waves in the sea over the frequency range from a few hertz to a few megahertz. The International System (SI) units are the pascal (Pa) for pressure (equal to one newton per square meter) and the watt per square meter (W/m^2) for sound intensity (the flow of energy through a unit area normal to the direction of wave propagation). In acoustics, it is more convenient to refer to pressures, which are usually much smaller than a pascal, and the consequent intensities with a different reference, the decibel. Intensity in decibels (dB) is ten times the logarithm to the base ten of the measured intensity divided by a reference intensity. The unit of reference intensity is the intensity of a plane wave that has a root-mean-square sound pressure of one micropascal (1 μPa). In the past, several other reference units have been used. *See* DECIBEL; SOUND INTENSITY; SOUND PRESSURE.

Wave propagation. The mathematical equation that sound obeys is known as the wave equation. Its derivation is based on the mathematical statements of Newton's second law for fluids (the Navier-Stokes equation), the equation of continuity (which essentially states that when a fluid is compressed, its mass is conserved), and a law of compression, relating a change of volume to a change in pressure. By the mathematical manipulation of these three equations, and the assumption that only very small physical changes in the fluid are taking place, it is possible to obtain a single differential equation that connects the acoustic pressure changes in time to those in space by a single quantity, the square of the sound speed (c), which is usually a slowly varying function of both space and time. This known as the wave equation [Eq. (3)], where p is the acoustic pressure; and

$$\nabla^2 p = c^{-2} \frac{\partial^2 p}{\partial t^2} \quad (3)$$

the differential operator ∇^2, known in mathematics as the laplacian, is, in many underwater acoustics applications, given in either spherical or cylindrical coordinates. *See* FLUID-FLOW PRINCIPLES; LAPLACIAN; NAVIER-STOKES EQUATION; WAVE EQUATION.

Knowing the sound speed as a function of space and time allows for the investigation of the spatial and temporal properties of sound, at least in principle. The mathematics used to find solutions to the wave equation are the same as those that are used in other fields of physics, such as optics, radar, and seismics. *See* WAVE MOTION.

In addition to knowing the speed of sound, it is necessary to know the location and nature of the sources of sound, the location and features of the sea surface, the depth to the sea floor, and, in many applications, the physical structure of the sea floor. It is not possible to know the sound speed throughout the water column or know the boundaries exactly. Thus the solutions to the wave equation are never exact representations of nature, but estimates, with an accuracy that depends on both the quality of the knowledge of the environment and the degree to which the mathematical or numerical solutions to the wave equation represent the actual physical situation.

There are a variety of analytical and numerical solutions to the wave equation that are useful. An important parameter in selecting the most appropriate solution is the ratio of acoustic wavelength to some of the characteristic dimensions of the environment.

The wavelength λ of a sound wave of frequency f is Eq. (4). Since the speed of sound is approximately

$$\lambda = c/f \tag{4}$$

1500 m/s (5000 ft/s), typical wavelengths for 100, 1000, 10,000 and 100,000 Hz are 15, 1.5, 0.15, and 0.015 m (50, 5, 0.5, and 0.05 ft), respectively.

If the wavelength is very small compared to the dimensions of oceanographic interest (such as depth, distance, sea floor, or sea surface roughness), good approximate solutions are those of geometric optics, known as ray theory. Rays that originate at the source follow paths that are determined by Snell's law. According to Snell's law, if a ray is launched from a source where the sound speed is c_0 at an angle θ_0 from the horizontal axis, the horizontal angle θ of the ray will, at any point along its path, obey Eq. (5),

$$\cos\theta/c = \cos\theta_0/c_0 = \text{a constant} \tag{5}$$

where c is the speed of sound at a point along the ray path. This statement allows ray paths to be determined for each angle θ_0 by either analytical or numerical computation. A simple example is shown in **Fig. 2**. As will be seen later, this limit allows for good insight into the spreading and the distribution of acoustic energy. *See* GEOMETRICAL OPTICS; REFRACTION OF WAVES.

The bending of rays is called refraction. The case shown in Fig. 2 is for a linear sound speed profile. It can be shown for this case that rays will follow circular paths. There will be a limiting ray, as shown, beyond which no ray will exist. The region beyond that ray is known as the shadow zone, and is well known to limit the detectability of objects by sonars. Other, more elaborate solutions to the wave equation are required when geometric distances become comparable to wavelength, such as a sound wave near a boundary or object, or when multiple waves are present. *See* SONAR.

The interaction of sound waves with the boundaries of the sea is essential for understanding underwater sound. By examining a highly idealized model, it is possible to gain a valuable understanding of some important properties of the interaction of acoustic fields with the sea surface and sea floor. It is useful to consider the case of a plane sound wave with a frequency f. The second-order derivative with respect to time on the right-hand side of the wave equation (3) yields for that term $c^{-2}\partial^2 p/\partial t^2 = -4\pi^2 f^2 c^{-2} p$. Substituting this into Eq. (3) gives Eq. (6), which is

$$(\nabla^2 + k^2)p = 0 \tag{6}$$

known as the Helmholtz equation. The quantity k is called the wave number, defined to be $k = 2\pi f/c$. If, as shown in **Fig. 3**, the acoustic wave is directed toward a boundary at an angle θ from the horizontal axis, x, the wave can be represented by Eq. (7),

$$p_i = A\sin(kx\cos\theta - ky\sin\theta - 2\pi ft) \tag{7}$$

where A is its acoustic pressure amplitude. There

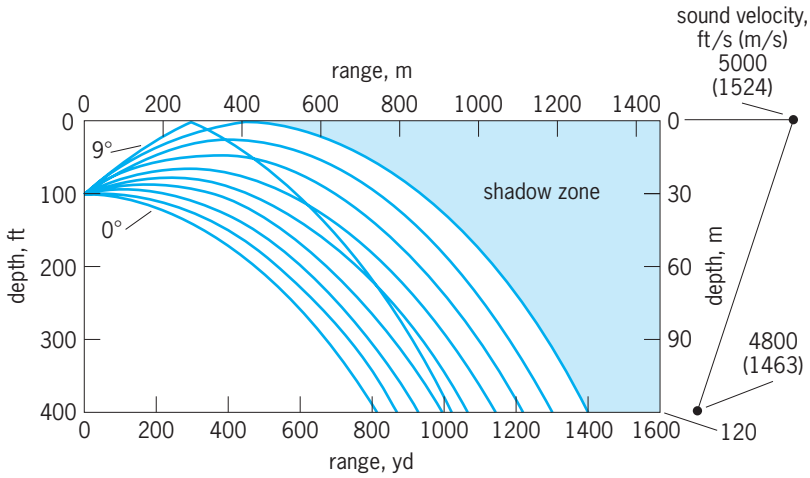


Fig. 2. Ray paths for a source near the surface of an ocean with a linear negative sound speed gradient, shown at the right. For this case, all rays follow the arcs of circles with different radii for each angle of emission. (*After R. J. Urick, Principles of Underwater Sound, 3d ed., McGraw-Hill, 1983, reprint, Peninsula Publishing, 1996*)

will also be a reflected wave, given by Eq. (8), whose

$$p_r = B\sin(kx\cos\theta + ky\sin\theta - 2\pi ft) \tag{8}$$

acoustic pressure amplitude is B. If the medium on the other side of the boundary is also a fluid, but with a sound speed of c_0 and a wave number k_0, a good "guess" at the solution for a wave traveling away from the boundary, as shown in Fig. 3, is given by Eq. (9), where D is the acoustic pressure amplitude.

$$p_t = D\sin(k_0 x\cos\theta_0 - k_0 y\sin\theta_0 - 2\pi ft) \tag{9}$$

At the boundary, $y = 0$, there are two conditions that have to be satisfied. The pressure on both sides has to be equal, and the vertical component of the fluid displacements has to be continuous, for all values of x. It is therefore required that $k\cos\theta = k_0\cos\theta_0$, or, remembering the definitions of k and k_0, $\cos\theta/c = \cos\theta_0/c_0$, which is identical with the Snell's law of Eq. (3). Returning to the conditions that have to be met at $y = 0$, from the continuity of pressure,

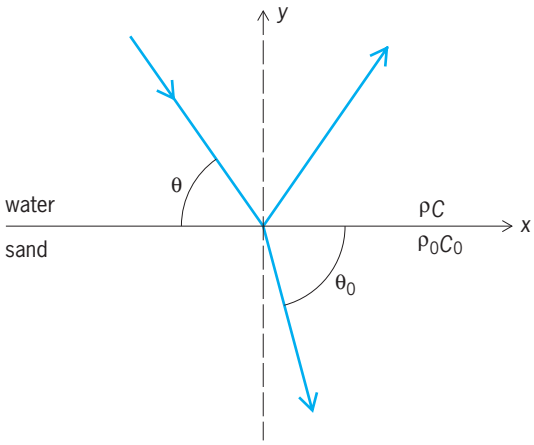


Fig. 3. Idealized model of reflection and transmission at a flat boundary of two fluids.

Eq. (10) must be satisfied.

$$A + B = D \tag{10}$$

A fluid's vertical velocity, u_y, can be shown to obey Eq. (11), where $\partial p/\partial y$ is the local vertical pressure

$$\frac{\rho \partial u_y}{\partial t} = \frac{\partial p}{\partial y} \tag{11}$$

gradient, and the density is ρ. In the upper fluid, at the boundary the total pressure is $p = p_i + p_r$. Taking the time derivative of u_y in both fluids to be equal at $y = 0$ leads to Eq. (12). Dividing Eq. (12) by Eq. (10) leads to Eq. (13) for B, where $Z = \rho_0 c_0/\rho c$.

$$\rho^{-1} k \sin\theta\{-A + B\} = -\rho_0^{-1} k_0 \sin\theta_0 D \tag{12}$$

$$B = \frac{A(Z\sin\theta - \sin\theta_0)}{Z\sin\theta + \sin\theta_0} \tag{13}$$

In the case of reflection from the air-sea surface, typical approximate values for the densities and sound speed for air and water, respectively, are 1.25 kg/m^3 (0.078 lb/ft^3) and 300 m/s (1000 ft/s), and 1000 kg/m^3 (62 lb/ft^3) and 1500 m/s (5000 ft/s). Therefore, Z is approximately 2.5×10^{-4}. It can also be shown from Snell's law that $\sin\theta_0$ will, for this case, always be a number close to 1. So the amplitude of the reflected wave will be almost exactly that of the incoming wave, only having the opposite sign. The sum of the amplitudes at the boundary of the incoming wave, p_i, and the reflected wave, p_r, is very close to zero. This boundary is called a pressure release boundary and is approximated as $p = 0$. Since $A \approx B$, it is also evident from Eq. (10) that very little energy is transmitted to the air.

A typical sandy sea floor has a sound speed approximately 10% higher and a density about twice that of seawater. From Snell's law it can be seen that at the angle $\theta = \arccos(c/c_0) \cong 25^\circ$ the direction of the refracted wave in the sea floor is parallel to the boundary. At that angle, it can be seen that $B = A$. That is, all of the acoustic energy is reflected. For angles $\theta < 25^\circ$, values of $\sin\theta_0$ become purely imaginary and B can be shown to be equal to A times a term that only represents a phase change. The term B/A is known as the reflection coefficient, and the angle $\theta_c = \arccos(c/c_0)$ is called the critical angle. This concept of total reflection is essential to the understanding of propagation in shallow water. It can be shown that the reflection is a minimum at normal incidence, implying that this is also the angle of maximum penetration into the sea floor and therefore important in the use of acoustics in the study of the internal structure of the seabed.

Although these two idealized examples do not represent the actual conditions of the sea, they do contain some of the physics that is observed. The sea surface is rarely flat, but it is always considered to be a pressure release surface. The sea floor is never a perfect fluid, and more realistic model extensions include absorption, viscosity, shear, and layering. The theories follow the same rules, namely, finding solutions in each medium and matching the boundary conditions of continuous pressure and displacement. *See* REFLECTION OF SOUND.

Deep-water propagation. During the middle years of the twentieth century, it was discovered that low-frequency sound, under some conditions, could propagate over very long distances with unusually small losses in intensity. Indeed, explosive signals were heard halfway around the Earth. The concept can be understood through the application of ray theory in deep water. Since the deep oceans are several thousand meters deep and major oceanographic features have dimensions many times larger than acoustic wavelengths, the use of geometric optics, or ray theory, is appropriate.

In many parts of the world, warm water overlays the deeper cold water. Sound speed near the surface can be dominated by the temperature, which is highest at or near the surface, and then decreases with depth until it becomes constant, in most oceans somewhere near 4°C (39°F). From Eq. (2) it can be seen that the effect of temperature is to decrease sound speed with depth, and the pressure term ($0.016Z$) causes an increase with depth. The sound speed profile as a function of depth (due to the temperature and depth) will have a minimum that is usually found to be at about 1000 m (3300 ft). An idealized picture is shown in **Fig. 4**.

For an impulsive omnidirectional sound source (typical of an explosive) that is located in depth at the minimum sound speed (Fig. 4*a*), rays are emitted in all directions and each will follow a trajectory

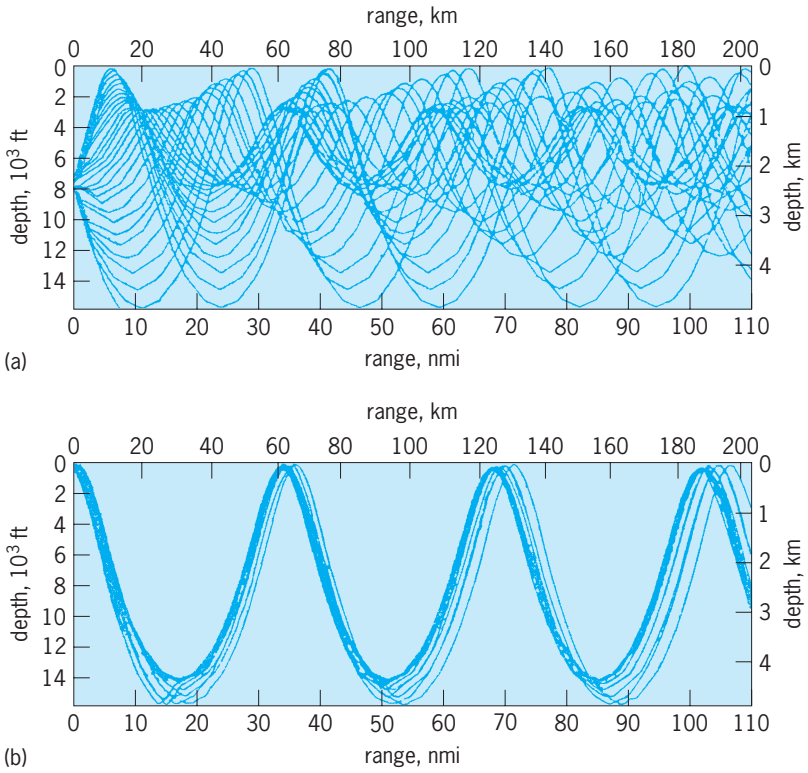


Fig. 4. Ray trace diagrams for deep-water propagation (*a*) for a source near the sound channel axis and (*b*) for a source near the surface—note the large region of shadow zones.

dictated by Snell's law. Those that are directed upward will gradually bend downward as they approach water with higher speeds. Those with steep enough trajectories will strike the sea surface. There will be a limiting ray that reaches the surface at zero degrees. Its angle of emission from the source, using Snell's law, is given by Eq. (14).

$$\theta_L = \arccos\left(c_{\min}/c_0\right) \tag{14}$$

All rays with smaller angles of emission will not reach the surface, and will bend toward the axis of minimum speed, intersecting the axis with the equal but opposite sign to that at which they were emitted. If the sound speed at the surface is less than that at the bottom, these rays will not reach the bottom as shown in Fig. 4. The same is true for the down-going rays. There will be a limiting ray with an angle again dictated by Snell's law, and rays that do not interact with either the surface or the sea floor and propagate outward from the source cylindrically rather than spherically.

The path that each ray will follow can be determined using Snell's law and the sound speed profile. (There are many computational methods available for this.) For a given range, it is possible to determine the family of rays that can be emitted from the source and pass through that point. It is also possible, knowing the ray paths, to compute the arrival time of each. It can be shown that the rays that arrive first are those that make the largest excursions from the channel axis, and the last to arrive is the ray that travels along the channel axis, which will be the most intense. The signal that is received on the channel axis from a distant explosive source begins with weak arrivals from the rays making the widest excursions and ends with a dramatic buildup that suddenly cuts off with the arrival of the ray that travels along the channel axis. The arrival of this last signal can give an accurate determination of the range from the source to the receiver, and is the basis for the SOFAR concept that was used for the location of downed aircraft in earlier times. *See* SOFAR.

For a sound source located near the surface, the rays that propagate without interaction with the surface or sea floor are more limited and leave large regions where sound does not penetrate (Fig. 4*b*). These shadow zones are vast, and it is only after large distances that the rays arrive near the surface. This sinuous behavior of the sound energy continues for very long distances with cyclic intervals of approximately 60–70 km (37–44 mi). The areas where there is focusing near the surface are called convergent zones. These are important for the detection of distant signals.

Of interest to the oceanographers and acousticians is the application of long-range, deep-water propagation to the study of ocean structure, such as warm and cold water eddies, the meander of ocean currents, internal waves, and global warming. For a fixed geometry of sources and receivers there will be specific ray paths, each migrating through the upper and lower depths. The arrival time of each ray is determined by the path and the sound speed along it. Experiments have shown remarkable temporal stability in many of the paths By observing the angular direction and arrival times, the paths of the arrivals can be identified. The measurement of small changes in the arrival times can be used to interpret oceanographic change. It is a form of tomography. *See* COMPUTERIZED TOMOGRAPHY.

There are many other mathematical solutions to the wave equation that are used in deep-water application. In order to obtain tractable solutions, many of these require assumptions about the homogeneity of the ocean. Of particular merit is the application of the parabolic approximation to long-range propagation for the case where there is a horizontal variation in oceanographic conditions. In most ocean applications, it is a valid to assume that the horizontal change in sound speed is sufficiently gradual to allow the range-dependent part of the Helmholtz equation to be approximated by a parabolic equation. The method is highly amenable to modern computation, and has been applied to many problems in acoustics, including shallow-water and atmospheric acoustics. *See* DIFFERENTIAL EQUATION.

Shallow-water propagation. Shallow water can be defined as the regions of the oceans associated with the continental shelves. Shallow depths are of the order of a few hundred meters or less. Unlike deep water, the sea floor plays a major role in the acoustic properties in shallow water. In shallow water, the application of higher frequencies is more suitable for both the detection of objects and for the study of some of the smaller-scale oceanographic processes. For lower frequencies, propagation over longer ranges requires the consideration of the many acoustic paths contributed by the continued interaction with the boundaries. Sound that reflects off the sea surface loses little energy. The reflections off the sea floor when the grazing angle is greater than critical are also nearly perfectly reflecting. Thus the sound field at longer ranges comprises many arrivals.

The principal method of solving the wave equation in shallow water is that of normal modes. The approach is to make the assumption that sound propagation in the ocean has no azimuthal dependence. Then the wave equation, or the Helmholtz equation, depends only on range and depth, and is separable into an equation of depth and one of range. The sound speed is assumed to have no range dependence, but it can depend on depth.

The range-dependent equation is the Bessel equation, whose solutions are well known. The vertical, or depth-dependent equation, is then solved. It will depend on the sound speed profile, the water depth, and the boundary conditions at the sea surface (the pressure release condition) and the sea floor. This leads to a set of solutions, called modes. Then, by representing the source, often as a point source at a depth z_0, the solution is obtained for any point in the water column by a sum of modes. Range dependence is difficult to accommodate to this method, but the normal-mode method is a very powerful

tool for understanding the complex way that signals travel through shallow water. Signals become distorted from their original form due to the changes in horizontal speed that result from the complex interaction of sound with the boundaries. The frequency dependence of sound speed is called dispersion. *See* BESSEL FUNCTIONS.

Ambient noise. A consequence of the remarkable transmission of sound is that unwanted sounds are transmitted just as efficiently. One of the ultimate limitations to the use of underwater sound is the ability to detect a signal above the noise. In the ocean, there are four distinct categories of ambient sound: biological, oceanographic physical processes, seismic, and anthropogenic. Seismic is in the frequency range called infrasound, and will not be covered. *See* INFRASOUND.

Biological sources. Biological sound is caused by both fish and marine mammals for the purpose of either communication or predation. The frequency range is broad. The sounds are sometimes very intense and geographically variable. At the low frequencies (those that would be called audible), the sounds of whales and other mammals dominate. Many of the species are migratory, giving rise to very high sound levels at certain times of the year at specific locations. There are currently intense studies of these creatures with the objective of assessing their numbers, their habits, and the purposes of their communications. At the higher frequencies, about 1–20 kHz, snapping shrimp dominate the sound in shallow water, especially around piers and other structures. From an engineering viewpoint, it is difficult to conceive of methods, other than avoidance, to process against these noises due to their high variability in space and time.

Physical processes. The natural physical causes of ambient noise are the sounds of breaking waves, surf, and rain noise.

There have been only a few studies of the physical mechanisms that create the sound. In the few measurements that have been made on individual breakers, there appears to be a wide variation in the frequency content. A plausible hypothesis is that the collapsing of the bubble cloud produced by the breaking wave generates a band of acoustic frequencies. This, however, has not been confirmed. Because winds are the general cause of breaking waves, this type of noise is referred to as wind-generated. It dominates the noise spectrum above about 500 Hz, but is a contribution down to at least 100 Hz. Wind-generated noise increases with wind speed, with the difference in level being about 30 dB between a very calm sea (sea state 1) and a violent sea (sea state 6). *See* OCEAN WAVES.

Anthropogenic sound. Shipping dominates the ambient sound in most oceans in the frequencies of about 10–150 Hz. The level clearly depends on the amount and type of shipping and the propagation conditions. The radiated noise from each class of ship is different, with the loudest being supertankers. Another anthropogenic sound, from oil prospecting, is seasonal and geographically limited, but the airguns and sparkers used have very high source levels.

Figure 5 shows an averaged compilation of typical sound levels that are due to anthropogenic and natural physical sounds. *See* ACOUSTIC NOISE.

Scattering and reverberation. The other source of unwanted sound is reverberation. Sound that is transmitted inevitably finds something to scatter from in the water column, at the sea surface, or at the sea floor. The scatter is usually in all directions, and some of it will return to the system that processes the return signals. Sources of scattering in the water column are fish, particulates, and physical inhomogeneities. The sea surface is, under normal sea conditions, agitated by winds and has the characteristic roughness associated with the prevailing

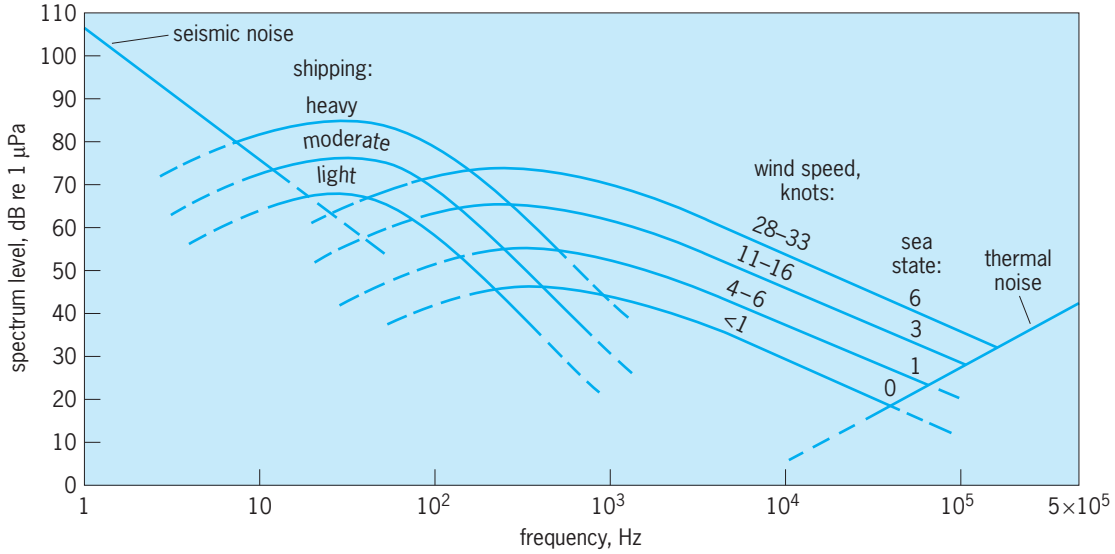


Fig. 5. Average values of ambient noise for both shipping and wind-generated sound, and representative values of seismic and thermal noise. 1 knot = 0.5 m/s. (*After R. J. Urick, Principles of Underwater Sound, 3d ed., McGraw-Hill, 1983, reprint, Peninsula Publishing, 1996*)

atmospheric conditions. Rough surfaces scatter sound with scattering strengths that depend on the roughness, the acoustic frequency (or wavelength), and the direction of the signal. The scattering is highly time-dependent, and needs to be studied with an appropriate statistical approach. The sea floor has inherent roughness and is usually inhomogeneous, both properties causing scatter. Although scatter degrades the performance of sonars, the characteristics of the return can be determined to enable its cancellation through signal processing or array design. Scattering can also be used to study the sea surface, the sea floor, fish types and distribution, and inhomogeneities in the water column. *See* SCATTERING LAYER.

There are still fundamental questions about the scattering of sound from the sea surface. At higher wind speeds, the classic theory of scattering from the sea surface underestimates the scattering strength by several decibels, enough to question either its validity or applicability. There is evidence that the layer of bubbles caused by the breaking of waves plays a significant role. Conclusive experimental proof of this hypothesis has not been obtained. The sea floor needs more study to determine the causes of scattering. That, however, may not yield a universal model because of the widely varying nature of the sea floor.

Experiments and ships. Both oceanography and ocean acoustics (underwater sound) have gained enormously from the developments in electronics, miniaturization, materials, and digital processing. It is now possible to place measuring instruments in the sea: on a buoy (either free-floating or tethered), on an anchored platform anywhere in the water column, or on underwater vehicles. The miniaturization of processing and data storage allows for the collection of large databases in real or near-real time. Communications now allow the dissemination of data to many locations at any distance. *See* INSTRUMENTED BUOYS; UNDERWATER VEHICLES.

Underwater acoustic instrumentation, whether for basic science studies, the detection of objects in the sea, or the exploitation of nature's resources, can be divided into active and passive types. Active measuring systems send and receive signals. Passive systems only receive signals. The active systems currently use everything from explosives and airguns, which create high-energy short pulses at characteristically low frequencies, to small hand-held systems with ultrahigh frequencies.

An acoustic signal is usually created by the dilatation of either magnetostrictive or piezoelectric materials. The design for their use is dictated by the desired energy, frequency, and directivity (beam width). The most common materials used for the reception of signals in both active and passive systems are piezoelectric. The receivers are commonly called hydrophones. Significant gains can be achieved in signal level by geometrically arranging sets of hydrophones or transmitting elements with the proper spacing. These are called arrays. Modern electronics and digital processing allow for many hydrophones to be placed in an array that can be towed from a ship, deployed horizontally, suspended horizontally from a stationary ship or platform, or suspended from an anchor. **Figure 6** depicts some of these concepts. Sources can be fixed or towed and, in some cases, mounted on platforms such as ship's sonars and depth finders. *See* ACOUSTIC SIGNAL PROCESSING; HYDROPHONE; MAGNETOSTRICTION; PARAMETRIC ARRAYS; PIEZOELECTRICITY.

The common requirement of all oceangoing research is a ship to carry equipment, scientists, and engineers to the site of an experiment, to deploy the instruments, maintain station, and to run on

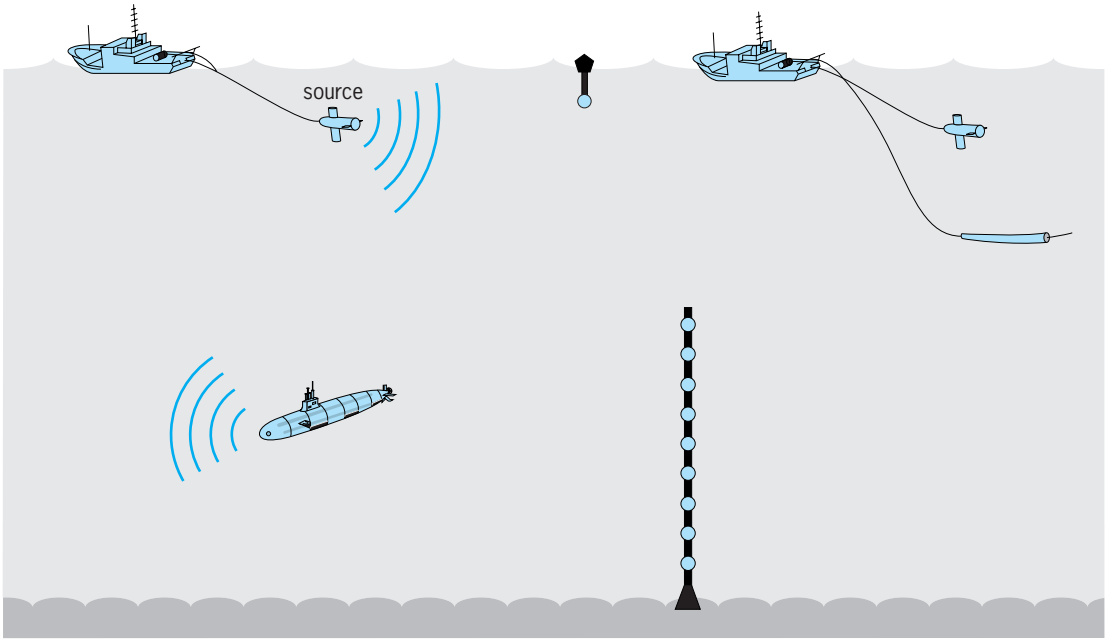


Fig. 6. Different sources and receivers used in deep-water and shallow-water experiments.

specifically determined routes. The planning and logistics for ocean experimentation are a vital part of the experience. The ships that are used for deep-sea work must be capable of withstanding storms and robust enough to be functional for the completion of experiments under difficult conditions. Nearshore operations usually do not have the same requirements. Often overlooked is the cost and tempo of working at sea. While ocean instrumentation has improved and its costs have dropped, the price of ships and the cost of operation have increased significantly. Although modern oceanographic vessels are much more capable than those of the past, they still sail at speeds of only a few knots. Seagoing is still a labor-intensive endeavor that cannot be overlooked in the planning of experiments. Sometimes aircraft are used for the deployment of sources and sensors. They are excellent within the limitations that are inherent in work with aircraft such as those on weight and volume. *See* OCEANOGRAPHIC VESSELS.

Ralph R. Goodman

Bibliography. L. Brekhovskikh and Yu. Lysanov, *Fundamentals of Ocean Acoustics*, 1982; C. S. Clay and H. Medwin, *Acoustical Oceanography*, 1977; C. S. Clay and H. Medwin, *Fundamentals of Acoustical Oceanography*, 1988; Y. Desaubes et al. (eds.), *Oceanographic And Geophysical Tomography*, 1990; G. Frisk, *Ocean and Seabed Acoustics*, 1994; B. R. Kerman (ed.), *Sea Surface Sound*, 1988; National Defense Research Committee, *Physics of Sound in the Sea*, 1945, reprint 1969 and 1989; I. S. Tolstoy and C. S. Clay, *Ocean Acoustics*, 1966, reprint 1987; R. J. Urick, *Ambient Noise in the Sea*, 1986; R. J. Urick, *Principles of Underwater Sound*, 3d. ed., 1983, reprint 1996; J. L. Worzel, M. Ewing, and C. L. Pekeris, *Propagation of Sound in the Ocean*, 1948, reprint 2000.

Underwater television

Any type of electronic camera that is located underwater in order to collect and display images. It must be packaged in a waterproof housing. The underwater camera may be packaged with its own recording device, or it can be attached to a television that is located on a ship, in a laboratory, or at a remote site. In the latter case, the images by the camera are called real-time images, as they are available at the same time that the camera is recording them. An underwater television may be used for sport, ocean exploration, industrial applications, or military purposes. Common imaged subjects are animals, coral reefs, underwater shipwrecks, and underwater structures such as piers, bridges, and offshore oil platforms.

During the daytime and at minimal depths, underwater television can be used to view objects illuminated with natural sunlight. For nighttime viewing or at very deep depths, artificial lights must be used. Light in the ocean is reduced in intensity quite severely: even in the clearest natural waters, a beam of blue or green light will be reduced in intensity by approximately 67% every 230 ft (70 m). Light that is propagating through an aqueous medium such as seawater or lake water will also be selectively reduced in intensity, based on wavelength. In most situations, the extreme reds and blues will be most severely attenuated, with the region of highest clarity being the yellow to blue-green wavelengths. *See* ELECTROMAGNETIC RADIATION.

Modern advances in television cameras have opened up a host of applications for underwater viewing, such as of standard-video, charge-injection-device, silicon-intensified-target, or charge-coupled-device cameras. These cameras can be adapted for high frame rate, low light level, or underwater color imaging. Computer modeling of underwater images can be used to predict the performance of an underwater imaging system in terms of range of viewing and quality of images as a function of the clarity of the water. The most spectacular use of an underwater television system was the discovery of the lost luxury liner *Titanic*. In order to find the *Titanic*, which was located at a depth of approximately 16,000 ft (4800 m), an underwater television system that consisted of a video camera equipped with lights was used. *See* CHARGE-COUPLED DEVICES; TELEVISION; TELEVISION CAMERA; UNDERWATER PHOTOGRAPHY.

Jules S. Jaffe

Bibliography. R. D. Ballard, The discovery of the *Titanic*, 1987; J. S. Jaffe, Computer modeling and the design of optimal underwater imaging systems, *IEEE J. Oceanic Eng.*, 15(2):101–111, 1990; R. C. Smith and K. S. Baker, Optical properties of the clearest natural waters (200–800 nm), *Appl. Optics*, 20:177–184, 1981.

Underwater vehicles

Submersible work platforms that can be operated either remotely or by an onboard crew. Unlike military submarines, which are essentially submerging ships, submersibles are not self-contained. All submersibles require some sort of surface-support vessel.

History

Perhaps the earliest underwater vehicles were the seventeenth-century ambient-pressure diving bells used mostly for recovering items from shipwrecks, in which the workers were in a dry environment but exposed to full depth pressure. Development of practical,crewed underwater vehicles began before World War I, with 1-atmosphere surface-tethered diving bells and armored diving suits. Both were intended to perform salvage tasks. Some of the bells could go to depths of several hundred feet and were equipped with external lights and a variety of tools.

Armored suits were one-person submersibles in the general shape of a diving suit. Depth capabilities to 700 ft (210 m) were possible since the operator/diver was not exposed to sea pressure. Bottom time at the work site was greatly increased over conventional diver capability. Problems with the flexibility and watertight integrity of the suits' articulating

joints limited use of these devices prior to World War II.

In 1930, the United States scientist William Beebe and engineer Otis Barton developed the *Bathysphere* ("deep sphere"), a thick-walled steel ball lowered into the sea on a steel cable. The interior provided sufficient room for two crew members and their equipment. Two quartz glass windows permitted outside viewing and photography. The cabin had a third window opening but that window was broken during installation and was covered with a steel plate. The *Bathysphere* made 32 dives from 1930 to 1934 at Bermuda, ending with a dive to 3028 ft (910 m). Beebe's classic book, *Half Mile Down*, tells the story of this pioneering deep-sea research program.

In 1950, an improved bathysphere, the *Benthoscope*, was developed by Barton. It was successfully tested off California to depths of 4500 ft (1350 m), but made few working dives. Japan developed the *Kuroshio* for oceanographic research in the early 1950s. Retired in the 1960s, it was capable of diving to 650 ft (200 m). This was the world's last operational bathysphere, although the diving bells of today are not much different in concept.

It was not until 1948 that the first practical untethered underwater vehicle was tested: the bathyscaph ("deep ship") *FNRS-2*, invented by Swiss physicist Auguste Piccard. It was an underwater free balloon, with buoyancy provided by lighter-than-water aviation gasoline contained in a large steel float (balloon). A thick-walled spherical cabin for two crew members was suspended beneath the float. The float moved vertically by adding weight (seawater) or losing weight by dropping steel ballast pellets. Small electric motors fitted with propellers provided limited horizontal and vertical movements.

From 1950 to 1978, the French Navy developed and operated bathyscaphs *FNRS-3* (1953) and *Archimede* (1963). In 1953, Piccard built his last bathyscaph, *Trieste*, which was sold to the U.S. Navy in 1958. In 1960, a Navy expedition took *Trieste* to a depth of 35,800 ft (10,700 m) at Challenger Deep near the island of Guam in the western Pacific, the deepest place in the world ocean. After the original *Trieste's* retirement in 1963, the Navy built two improved bathyscaphs called *Trieste II*. Both had a maximum depth capability of 20,000 ft (6000 m).

Archimede was retired in 1978. It was the last crewed submersible in the world capable of diving to the deepest place in the sea. Four years later, *Trieste II* was retired; it was the last operating bathyscaph in the world. This type of deep submersible vehicle had served 35 years; much of this time, it was the only way to work in the deep ocean.

These retirements were not the end of deep-diving submersibles. New construction materials and technologies have provided the means for building submersibles that are much smaller, less costly to operate, and safer than the bathyscaphs.

During the 1960s and early 1970s, divers and crewed submersibles dominated the underwater scene. However, new types of vehicles were under development in research laboratories during this time. A major development was the remotely operated vehicle (ROV), a tethered vehicle controlled from the surface of the sea. The autonomous underwater vehicle (AUV), essentially an underwater robot, was also under active development but its maturity would take much longer than the ROVs.

Almost all early submersible designs, technologies, and operational techniques were developed by navies, the U.S. Navy in particular. Military missions and related science needs were the primary forces driving research on underwater operations during this period. By the mid-1970s, civil applications were stimulated by the steep oil price increases worldwide. This led to increased offshore oil and gas exploration and production activity. Opening new resource areas, such as the North Sea, accelerated development of better tools and techniques for conventional deep-sea divers. Beginning in 2004, this pattern of technological demand began repeating itself as oil prices soared again. Exploration and production now go to depths as great as 2 mi (3 km), where only submersibles can do the undersea work. *See* DIVING; OIL AND GAS, OFFSHORE.

Types

Underwater vehicles can be grouped into three general categories: crewed deep submersible vehicles (DSVs), tethered remotely operated vehicles (ROVs), and untethered autonomous underwater vehicles (AUVs). In recent years, the term "human occupied vehicles" (HOV) has sometimes been substituted for DSV.

There are also hybrid vehicles, which combine two or three categories on board a single platform. Within each category of submersible, there are specially adapted vehicles for specific work tasks. These can be purpose-built or modifications of standard submersibles.

Crewed deep submersibles. Since the late 1950s, over 200 crewed submersibles have been put into service worldwide. By 2005, about 75 were operational; the majority support tourist operations, with marine science applications being the second greatest. There are five types of these vehicles: 1-atm untethered vehicles; 1-atm tethered vehicles, including observation/work bells; atmospheric diving suits; and diver lockout vehicles. While they differ mainly in configuration, source of power, and number of crew members, all carry a crew at 1-atm (10^2-kilopascal) pressure within a dry, pressure hull. The fifth type of crewed submersible is the wet submersible, in which the crew is exposed to full depth pressure.

The purpose of the DSV is to put the trained mind and eye to work inside the ocean, often referred to as in situ work. The earliest submersibles had very small viewing ports fitted into thick-walled steel hulls. In the mid-1960s, experimental work began on use of massive plastics (acrylics) as pressure hull materials. Today, submersibles with depth capabilities to 3300 ft (1000 m) are being manufactured with pressure hulls made entirely of acrylic. Essentially the hull is now one huge window.

Fig. 1. Two-person deep submersible vehicle *Deep Rover* with all-acrylic pressure hull. (*Deep Ocean Engineering, Inc.*)

1-atmosphere untethered vehicles. These are submersibles that carry 1–64 occupants at 1-atm pressure (**Fig. 1**). The submersible is battery-powered (though fuel cells have been tested), free-swimming, and capable of limited horizontal travel across the seabed.

The four deepest-diving crewed submersibles are operated by France (1), Russia (2), and Japan (1). Three can dive to 20,000 ft (6000 m), and Japan's *Shinkai 6500* can go to 21,323 ft (6500 m). Though the deepest place in the ocean is about 36,000 ft (11,000 m), there is no crewed vehicle at present that can reach this depth. However, a submersible designed to operate at 20,000 ft can reach 98% of the sea floor worldwide. Therefore, designing for two-thirds the maximum depth of the ocean is a reasonable engineering and cost goal.

In 2005, two programs began constructing new crewed submersibles for marine science work. At the Woods Hole Oceanographic Institution, a 21,323-ft (6500-m) vehicle is being designed to replace the 13,000-ft (4000-m) *Alvin*, built in 1964, and in the Peoples Republic of China a 22,460-ft (7,000-m) crewed submersible is being built.

Observation/work bells and 1-atm tethered vehicles. These carry a crew of two or three at 1-atm pressure. Power is supplied by batteries, or from the surface through an umbilical cable. The vehicles (**Fig. 2**) always operate with a lift cable from the surface and are designed for panoramic viewing and a high degree of maneuverability and station-keeping, and are fitted with manipulators (external mechanical arms and hands) for work tasks. Many of these tethered vehicles are capable of limited maneuvering in three dimensions. They are primarily designed for performing manipulative work around underwater structures mostly in support of offshore oil and gas development. Maximum depth is 2300 ft (700 m) [**Fig. 3**].

Atmospheric diving suits (ADS). These are designed for a one-person crew at atmospheric pressure. In operation, they are similar to the 1-atm tethered vehicles. In this case, the vehicle is anthropomorphically configured (**Fig. 4**) and power for maneuvering is manual or electrical via a surface umbilical cable. The suit offers a high degree of manipulative dexterity and maneuverability. Atmospheric diving suits derive propulsive power from thrusters or manually by the operator's walking. The maximum depth of operation is 2000 ft (600 m).

Diver lockout submersibles. These can be either part of a free-swimming vehicle or a tethered bell system. They carry both an operating crew and a dive team. One section of the vehicle's pressure hull is at 1-atm pressure, while the other section can be pressurized to ambient pressure to permit diver egress. Power may be derived from batteries, advanced air-independent energy systems (such as fuel cells), or a surface-connected umbilical cable. Diver-breathing gases and heating may be carried aboard or supplied through an umbilical cable. A maximum operating depth (for diver lockout) is about 820 ft (250 m). Only a few diver lockout vehicles still exist and they are rarely used for commercial work. Navies use free-swimming ones for special warfare missions.

The more common way of delivering commercial divers to work sites is to use a cable-lowered diving

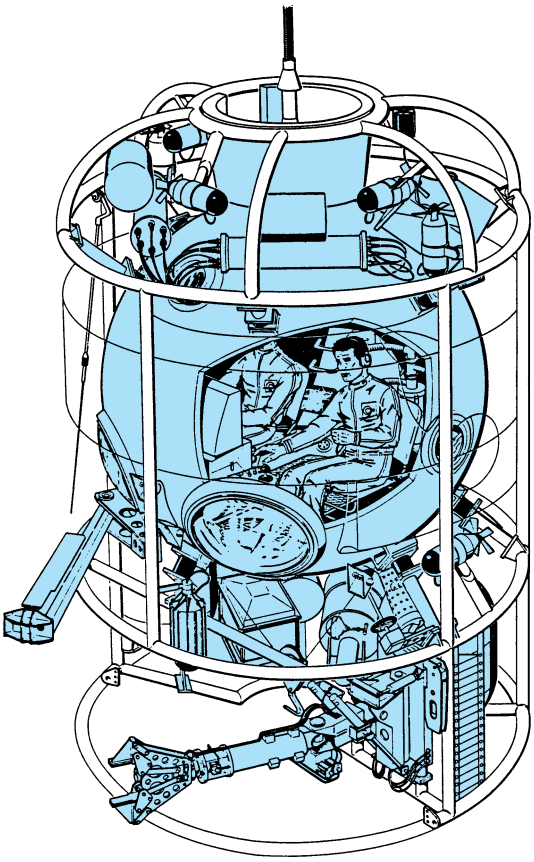


Fig. 2. Observation/work bell. (*Comex Industries*)

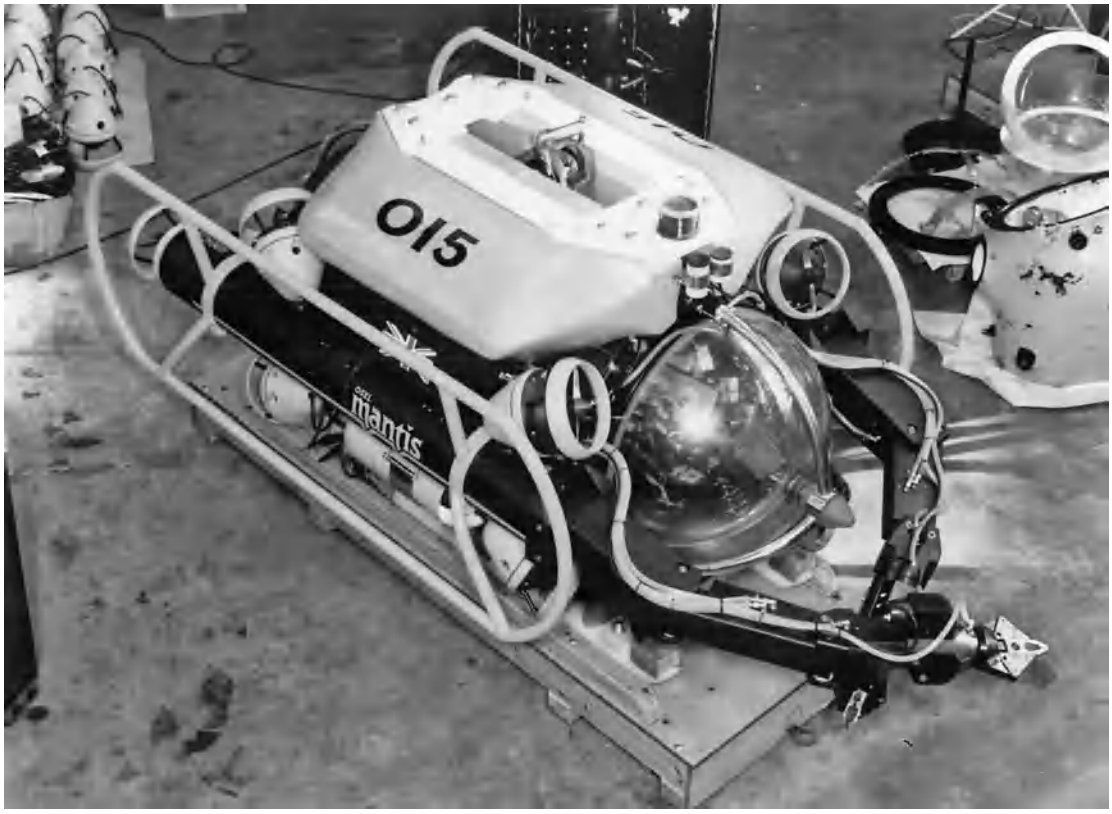


Fig. 3. One-atmosphere tethered vehicle *Mantis* (*OSEL Group, Ltd.*)

Fig. 4. HARDSUIT atmospheric diving suit. (*Oceanworks International*)

bell called a personnel transfer capsule (PTC). The divers are delivered and recovered from the dive site in a dry bell pressurized to equal the water pressure at the dive depth. A hatch in the bottom provides a means for the divers to leave and reenter the bell. On returning to the surface, the bell is mated to a deck decompression chamber (DDC) on aboard the support vessel, where the divers can decompress over periods ranging from hours to days.

Wet submersibles. There are two types of wet submersible vehicles: simple diver propulsion units in which a diver holds onto and steers through the water, and swimmer delivery vehicles (SDVs) in which the diver either rides inside (**Fig. 5**) or sits astride. In all cases, the personnel are exposed to full ambient pressure and water temperature. These vehicles are battery-powered and designed for shallow depths, not much more than 100 ft (30 m). The breathing air is carried in on-board tanks or on the diver's back. A few swimmer delivery vehicles are designed so that the interior of the cabin does not completely flood. The crew sits in a trapped air bubble, keeping the upper halves of their bodies in pressurized air. Facemasks or regulators are not necessary while passengers remain in the vehicle. While most of these vehicles have been developed for recreation, navy special warfare personnel also use similar vehicles.

Remotely operated vehicles. The first ROVs were developed in the late 1950s for naval use. By the mid-1970s, they were being widely adapted for the civil

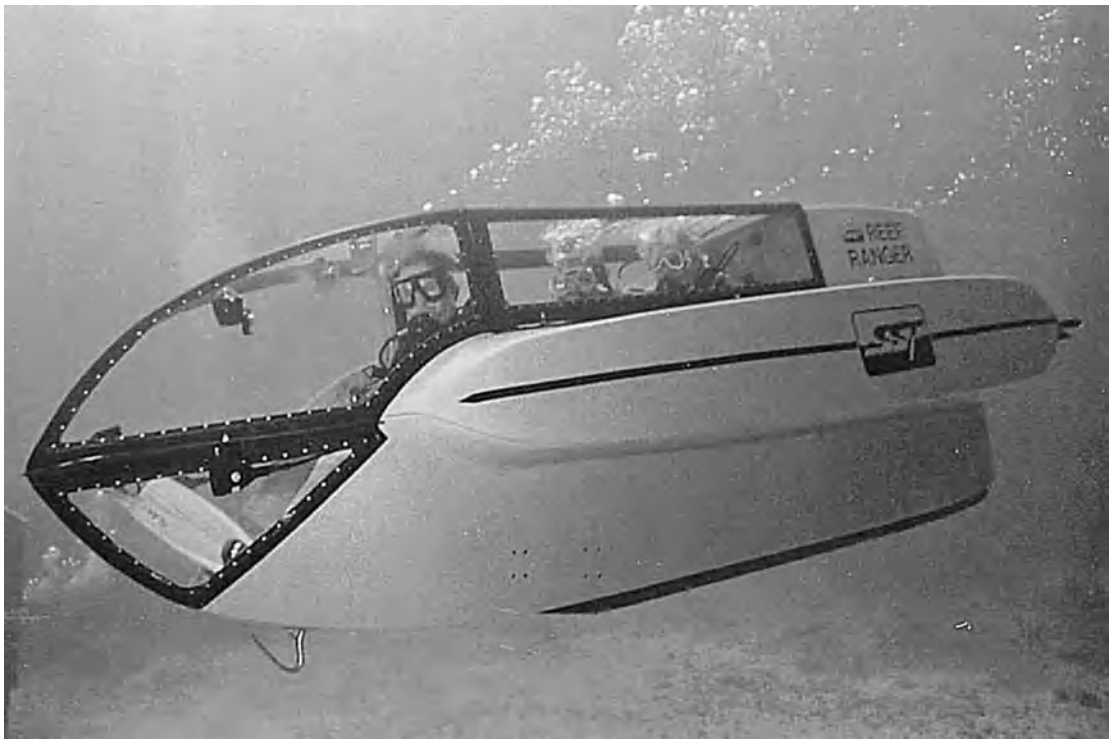


Fig. 5. Reef Ranger wet submersible. (*Submersible Systems Technology*)

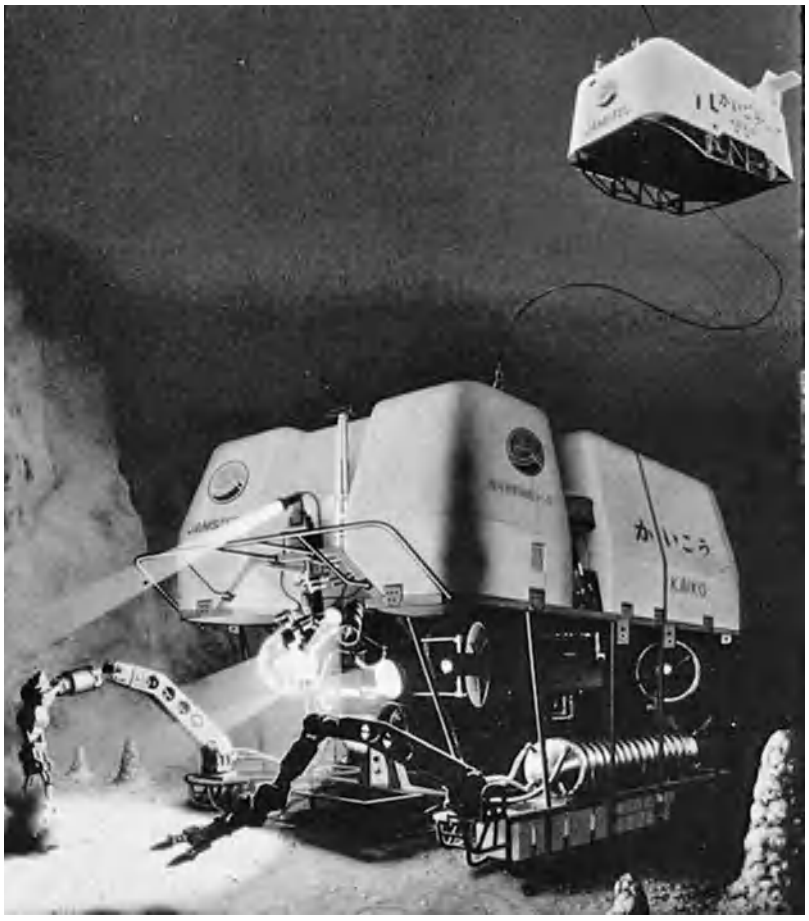


Fig. 6. Japan's 36,000-ft (11,000-m) *Kaiko*, a large working ROV. (*Jamstec & Mitsui Engineering and Shipbuilding*)

sector. In the offshore oil and gas industry, the ROV has substantially replaced both crewed submersibles and divers.

The rapid acceptance of these submersibles is due to their relatively low cost and the fact that they do not put human life at risk when undertaking hazardous missions. However, their most important attribute is that they are less complex. By virtue of their surface-connecting umbilical cable, they can operate almost indefinitely since there is no human inside requiring life support and no batteries to be recharged. There are four types of ROV: tethered free-swimming, towed, bottom-reliant, and structurally reliant.

Tethered free-swimming vehicles. These are the dominant remotely operated vehicles, and nearly 5500 have been constructed since the early 1970s. This type of ROV can be divided into two use categories: inspection vehicles (essentially swimming cameras) and work vehicles (equipped with a variety of tools and sampling devices). Costs of ROVs range from about $15,000 to several million dollars. *See* UNDERWATER PHOTOGRAPHY.

The vehicle itself is only part of an operating system, consisting of a control/display console, a power supply, a cable winch with a cable/vehicle handling apparatus, an umbilical cable (which transmits power and control commands down to the vehicle and data and television signals back to the surface), and the vehicle handling cage or "garage" (which is optional). The ROV is always equipped with a television camera, lights, and thrusters that provide three-dimensional maneuverability. In addition, there can be depth sensors, a wide array of manipulative and acoustic devices, and special

instrumentation to perform a variety of work tasks. For complex or deep-diving operations, the submersible is often equipped with an acoustic tracking system and a high-resolution sonar. This permits the surface support vessel to maintain close control of the vehicle relative to other targets on the sea floor. These vehicles can range from the size of a personal computer to that of a compact automobile. Larger ROVs are configured as basic work platforms that can use a wide variety of optional equipment according to the mission. *See* SONAR; UNDERWATER SOUND; UNDERWATER TELEVISION.

In the late 1980s, a few remotely operated vehicles with depth capabilities to 20,000 ft (6000 m) were put into service for deep ocean inspection and object recovery missions. In March 1995, the $50 million Japanese ROV *Kaiko* dove to 35,798 ft (about 11,000 m) into the Challenger Deep area off the Marianas Islands, equaling the 1960 record of the U.S. Navy's crewed Bathyscaph *Trieste*. Subsequently *Kaiko* returned to this site twice, demonstrating there are no depth limits to ROV capabilities. *Kaiko* was lost on a 6000 m dive in 2004; however, the Japanese government plans to replace it (**Fig 6**).

The vast majority of these vehicles have been operated at depths of less than 2000 ft (600 m). However, offshore oil and gas is being produced at depths greater than 3000 ft (1000 m), well beyond diver depths. Exploratory drilling occurs at 10,000 ft (3000 m) depth and it will not be long before production begins at these depths. Here, remotely operated vehicles are not an option but a necessity and several have been developed for these applications. Long-penetration vehicles are being developed to inspect the insides of several-mile-long aqueducts, ocean outfalls, and other large pipelines. For example, in 1998, a special ROV, *Phoenix*, made a record-setting 6-mi-long, (10,000-m) tunnel inspection in Finland.

Towed vehicles. Often called "sleds," these are usually equipped with instruments and imaging systems for long-range surveying, search, and reconnaissance (**Fig. 7**). As they are towed by a surface ship, their depth is controlled by adjusting the length of tow cable. Some sleds have fins to stabilize their underwater flight. There are two varieties within this class of vehicle. One type is towed above the bottom and is instrumented for scientific research or broad-area reconnaissance. The maximum depth of this type of vehicle is 20,000 ft (6000 m), though there are only a few in the world that can operate this deep. The second type is designed to be towed while in contact with the bottom to lay and bury pipelines or cables. These are basically specialized sea-floor plows.

Bottom-reliant vehicles. These are generally large, massive vehicles designed for a single-purpose work such as pipeline and cable burial, pipeline inspection, bottom excavation, and trench backfilling. They are controlled by an operator team on board a dedicated support ship, and maneuver on the bottom using

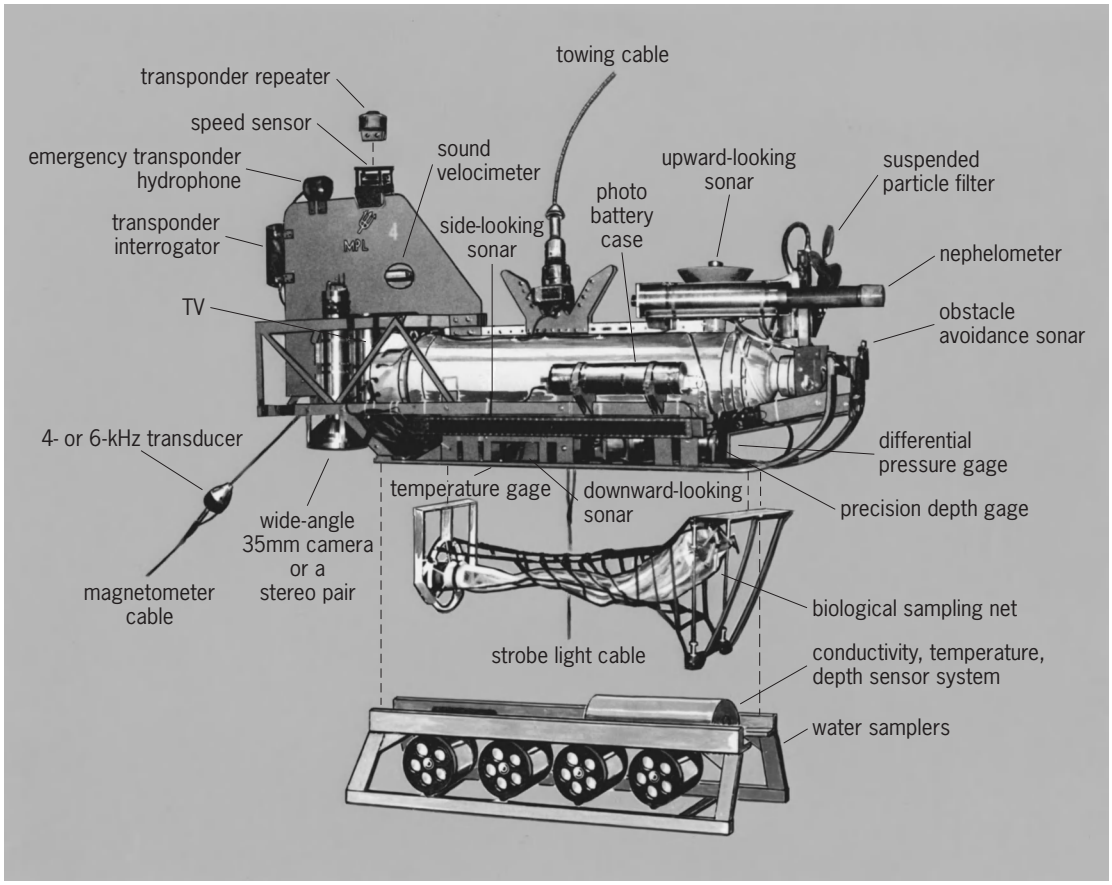


Fig. 7. Components of a towed vehicle. (*Marine Physical Laboratory, Scripps Institute of Oceanography*)

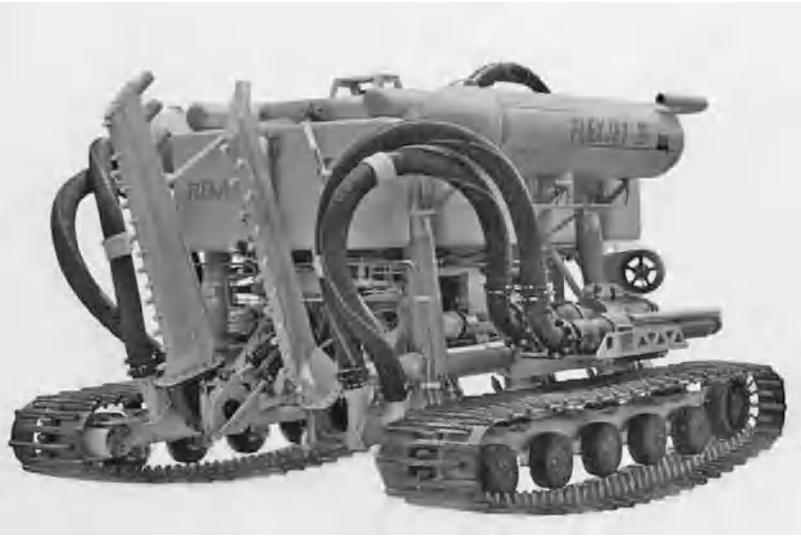


Fig. 8. Bottom-reliant vehicle. (*Sonsub International*)

wheels, caterpillar-like tracks, or Archimedes screws (**Fig. 8**).

Structurally reliant vehicles. These vehicles obtain power and control from the surface, but are designed to operate in contact with a fixed structure such as an oil platform. They can run on wheels or be propelled by push-pull rams as they crawl along a pipeline or other structure (**Fig. 9**). These vehicles are purpose-designed and conduct such tasks as ship hull and structure cleaning, inspection and maintenance of subsea production systems, and in-water repair tasks.

Autonomous underwater vehicles. These are crewless and untethered submersibles which operate independent of direct human control. Practically, AUVs are mobile instrumentation platforms with propulsion, powered by on-board batteries or fuel cells. A preprogrammed, on-board computer controls their operations, although a few experimental vehicles have been controlled by acoustic commands from the surface. The data they collect is stored and often processed on-board and, after surfacing, transmitted to other stations via satellite or radio link for further processing. Autonomous underwater vehicles' depth capability is from a few hundred feet to over 20,000 ft (6,000 m).

Fig. 9. Structurally reliant vehicle used for servicing seafloor oil well structures. (*Kvaener Brug A/S*)

First developed by the military in the early 1970s, numerous operational AUVs have entered the civil sector in recent years. Military applications include such tasks as minefield location, mapping, and neutralization; minefield installation; submarine decoys; military oceanography; and covert intelligence collection. Several navies now have operational AUV programs which actively use these vehicles in fleet operations.

Civilian tasks include site monitoring, basic oceanographic data gathering, under-ice mapping, offshore structure and pipeline inspection, and bottom mapping. Autonomous submersibles are particularly useful where human presence is not required and for long-duration operations.

Autonomous underwater vehicles span a wide range of sizes and capabilities, related to their intended missions. Large, transport-class platforms, about 33 ft (10 m) in length and 11 tons (10 metric tons), have been designed for missions requiring long endurance, high speed, large payloads, or high-power sensors. Smaller, network-class platforms, about 3.3 ft (1 m) in length and weighing 220 lb (100 kg), address missions requiring portability, multiple platforms, adaptive spatial sampling, and sustained presence in a specific region. Historically, most vehicles have been propeller-driven, but several new, variable-buoyancy, underwater gliders are emerging in the network class.

Over 100 autonomous underwater vehicles have been designed and built worldwide since 1963 by government laboratories, industry, and academia. Through the mid-1990s, most AUVs in all classes were research prototypes. By 2000, an increasing number of operational military and commercial vehicles were available, particularly in the network class. Developments in microprocessors, memory, batteries, and fuel cells are leading to higher performance in smaller platforms. Likewise, fiber optics and micro-electro-mechanical systems are spawning a new generation of compact, high-precision, low-power sensors ideally suited for payloads on autonomous underwater vehicles. The Global Positioning System (GPS) and emerging global communication services are enabling timely access to data and real-time control. *See* MICRO-ELECTRO-MECHANICAL SYSTEMS (MEMS); SATELLITE NAVIGATION SYSTEMS.

A major obstacle to more extensive use of AUVs is the limit on the amount of on-board power they can carry, thus limiting mission duration. Most submersibles of this type are powered by batteries. True long-duration missions will require an alternative to battery systems. Research is ongoing to find other power systems, such as fuel cells and solar cells which can be mounted on the vehicle. *See* BATTERY; SOLAR CELL.

Transport class. In the transport class, the *Mobile Undersea Systems Test Laboratory* is the world's

largest autonomous underwater vehicle: 33 ft (10 m) long, 4 ft (1.2 m) in diameter, weighing 10 tons (9.2 metric tons), rated to 1970 ft (600 m) depth, with an 87-mi (140-km) range and 1.1-ton (1-metric-ton) payload capacity. Other transport-class vehicles rated to full ocean depth include the French *Epaulard*, the Russian *MT-88*, the Chinese *CR-01*, and the *Theseus*, developed for the Canadian Department of National Defense. *Theseus* is rated to 3280 ft (1000 m) depth with an operating range of 466 mi (750 km) and an endurance of 100 h. In 1996, *Theseus* completed two missions under the Arctic icecap of 218 mi (350 km) and 50 h in duration.

Intermediate class. Intermediate between transport class and network class are vehicles such as the Marine Utility Vehicle System (MARIUS). At 14.8 ft (4.5 m) long, 2 × 3.6 ft (0.6 × 1.1 m) in rectangular cross section, and rated to 1970 ft (600 m) depth, MARIUS has an operating range of 31 mi (50 km) and is used for seabed inspections and environmental surveys in coastal waters. The Norwegian-built *Hugin 3000* AUVs have been quite successful commercially in doing seafloor surveys to depths as great as 9000 ft (3000 m) and for missions as long as 48 h.

Network class. Typical high-performance, propeller-driven vehicles in the network class include the Autonomous Benthic Explorer (ABE) and the Remote Environmental Monitoring Units (REMUS). Developed by the Woods Hole Oceanographic Institution, ABE is 992 lb (450 kg) in air, rated to 3.7 mi (6000 m) depth, and has an operating range of 62 mi (100 km). In 1997, ABE mapped magnetic anomalies associated with deep-ocean spreading zones in the North Pacific. REMUS, a highly portable, shallow-water vehicle is 5 ft (1.5 m) long, 8 in. (20 cm) in diameter, 88 lb (40 kg) weight in air, with up to a 50-mi (80-km) range (**Fig. 10**). REMUS has mapped water properties and performed mine reconnaissance to the edge of the surf zone. The U.S. Navy has used one of these submersibles for mine countermeasure work in the Arabian Gulf.

Network-class, buoyancy-driven vehicles are small, near-neutrally-buoyant platforms that move vertically and horizontally through the water because of small changes in buoyancy. One such vehicle, the Virtual Mooring Glider (VIRMOG), is fitted with battery-operated buoyancy controllers, and can operate with an estimated 40-day, 497-mi (800-km) range independent of temperature gradient. VIRMOG is designed for both deep- and shallow-ocean sampling, and uses changes in buoyancy for propulsion (underwater glider). Batteries are used as the energy source for the variable-buoyancy engine. One mission envisioned for VIRMOG is as a holding station at a fixed location, thus virtually emulating a mooring.

Hybrid vehicles. Hybrid vehicles are those that combine crewed vehicles, remotely operated vehicles, and divers. For example, the hybrid *DUPLUS II* can operate either as a tethered free-swimming ROV or as a 1-atm tethered crewed vehicle. This evolved to provide capability for remotely conducting those tasks for which human skills are not needed, and then to put the human at the place where those skills are required. Other hybrid examples include ROVs that can be controlled remotely from the surface or at the work site by a diver performing maintenance and repair tasks, bottom-crawling ROVs controlled by a diver to anchor a pipeline or cut a trench, and small "scout" ROVs that can "fly off" from a larger ROV. The Woods Hole Oceanographic Institution's *Argo-Jason* is an example of the fly-off system. It combines a towed vehicle (*Argo*) with a tethered free-swimming vehicle (*Jason*). *Argo* conducts large-area search or reconnaissance, while *Jason* (carried within and deployed from *Argo*) conducts detailed inspection and manipulation of objects of interest encountered during the towing phase.

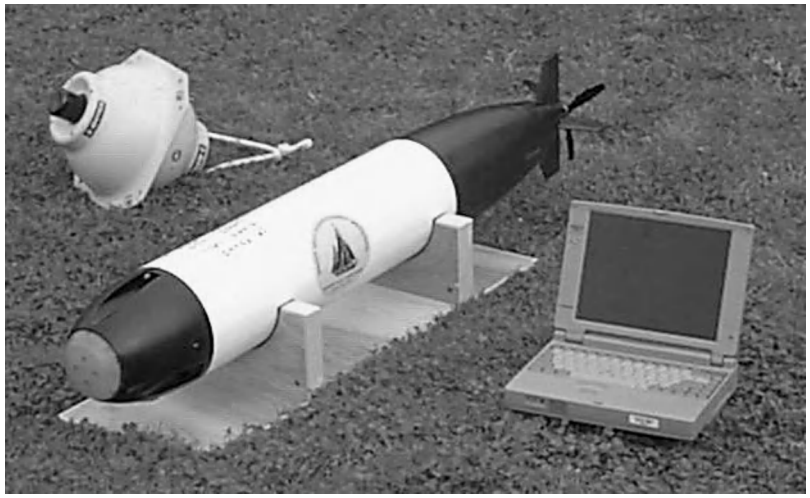


Fig. 10. REMUS is rated for continental shelf depths (355 ft; 200 m) and is compact and lightweight (40 kg; 88 lb). It is depicted with its laptop computer for downloading missions and uploading data, and one of the acoustic transponders used for navigation. (*Woods Hole Oceanographic Institution*)

Applications

Almost every conceivable underwater work task from military missions to recreational activities and from offshore oil and gas to environmental conservation has used submersibles. While there will always be a need for divers, as working depths increase the need for submersibles will continue to increase. The question of whether crewed submersibles will be replaced by ROVs and ROVs by AUVs is not valid. It is best to consider these types of vehicles, and all their variants, as a "family" of tools. The in situ worker chooses the best ones for a given job. A sample of work areas are listed below.

Offshore oil and gas activities. At one time, there was a clear distinction between the tasks that remotely operated vehicles performed in support of offshore oil and gas operations and those performed by divers and crewed submersibles. Over time, the capabilities of ROVs have increased to the point that almost every task once done by divers and crewed submersibles can now be done with ROVs, often better, and always at lower cost. As a result, crewed submersibles are rarely used in this business.

Observation, and video or photographic documentation. This is of the strongest areas of ROV uses, although in a

few cases some crewed vehicles are still used. The latter tend to be the 1-atm tethered and atmospheric-diving-suit (ADS) type submersibles. In shallow depths (<115 ft or 35 m) the diver is preferred. Observational tasks include determining the geometry and position of pipelines and cables following installation, determining the condition of a pipeline's concrete coating after installation, obtaining accurate positions of pipeline tie-in locations, leak detection monitoring, inspecting a wellhead for structural integrity, investigating accidents, and accumulating information needed to develop salvage or retrieval plans. Also included are bottom surveys such as pipeline and cable route selection, verification surveys, hardware site installation surveys, and debris identification and location mapping.

Structural nondestructive testing. This technique is used to locate and identify bent, broken, or missing structural members, identify and map debris location on structures, assess and remove marine growth, locate suspended pipeline and cable sections, perform structural-member thickness measurements, detect and measure cracks and flaws, and measure the effectiveness of the structural cathodic protection system.

Support of diver operations. Remotely operated vehicles can provide diver support and monitoring, including initial diving-gear checkout for leaks, continuous diver monitoring of job performance, safety, assistance in locating the diver's position, topside understanding of diving conditions, evaluation of the dive site in terms of diver safety prior to dive, provision of an additional and mobile light source, monitoring and inspection of the diver's work, and television or photographic documentation of the diver's work.

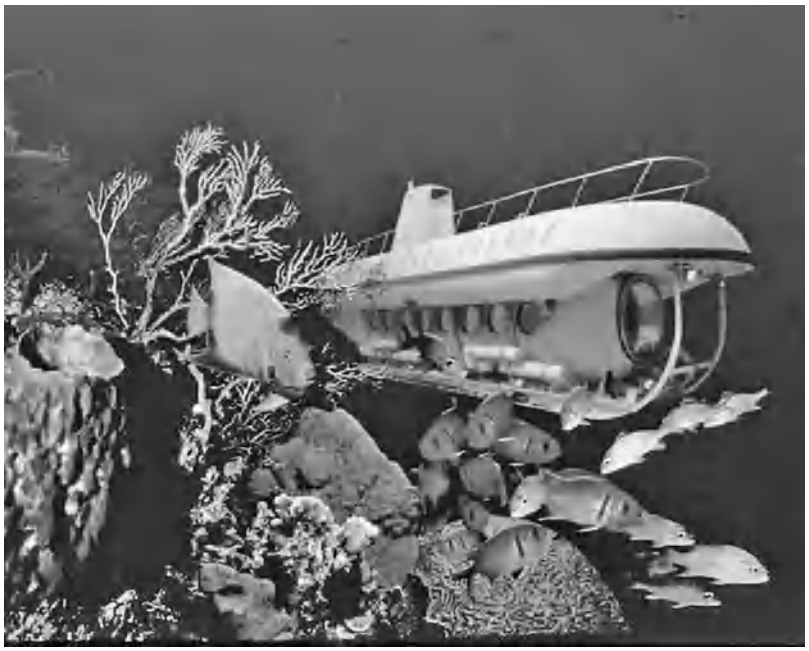


Fig. 11. Atlantis 48-passenger tourist submarine. (*Atlantis Submarines International*)

Object search, location, and retrieval assistance. Underwater vehicles are used to locate lost or abandoned objects, to determine the object's position relative to the surface salvage or retrieval vessel, and to provide assistance in attaching lift lines or cables to objects for retrieval. The majority of underwater vehicles cannot lift items weighing more than 110–220 lb (50–100 kg). Often separate lift systems (underwater elevators) are employed to work together with the submersible to bring heavy loads to the surface.

Oversight of underwater activities. Underwater vehicle are used to monitor underwater construction from surface platforms (such as ships, barges, and docks). These tasks include grouting operations, piling installation, alignment and orientation measurements during structural installation, observation of pipeline pull-in procedures, pipeline weighting and backfilling operations, and touchdown operations (for example, pipeline installations or spudding-in of drills).

Drilling support assistance. This is supplied by locating and retrieving objects, debris, and equipment; removing sediment to permit observation of components; connecting and disconnecting shackles and hydraulic/electric lines; collecting grout and bottom samples; operating valves; observing the drillstring operation and blowout preventer installation; and placing and recovering acoustic marker beacons.

Support of routine deepwater production operations. With hydrocarbon production occurring in increasingly deep waters that are well beyond diver depths, the routine management, inspection, and maintenance of seafloor structures with their associated pipelines is being done with ROVs. As the AUV technologies and techniques mature, eventually they will be used for tasks where a tethered vehicle is not feasible. In all cases, the sea-floor structures will be configured to be vehicle friendly for the mechanical arms (manipulators) and electronic eyes on the submersibles.

Marine leisure. This sector is the single largest user of crewed submersibles. Built in the early 1960s, the *Auguste Piccard* was the first tourist submarine. This 64-passenger submersible was used at the 1964 Swiss National Fair in Geneva. Crewed by pilot, copilot, and stewardess, it carried over 34,000 paying passengers during the 14-month-long fair.

In 1985, a Canadian company, Atlantis Submarines, built the 28-passenger *Atlantis I*, which was placed into service at Grand Cayman Island. This began a business sector that has employed some 60 crewed submersibles since that time. While a few of the early tourist submarines were conversions of existing DSVs, 51 were built for the business sector. The largest is the 66-passenger *Atlantis XIV* operating off Waikiki Beach, Hawaii. Thirteen *Atlantis* submarines are now in operation at Grand Cayman, Barbados; St. Thomas, U.S. Virgin Islands; Aruba; Cozumel; Hawaii; and Guam (**Fig. 11**). Tourist submersibles have operated in 34 countries throughout

the world. By 2005, they had carried nearly 12 million passengers with no serious incidents. A typical dive lasts about 1 h and goes to depths of 100–150 ft (30–46 m). There are also two small converted commercial-type submersibles at Grand Cayman that take two passengers to depths of 1000 ft (300 m).

The large passenger-carrying submersible business is a mature market sector. The best sites are occupied and few new submarines will be added to the fleet in the future. There is a new development direction, the resort-sized submersible for locations that cannot support a 40–50-passenger vehicle. These DSVs have a 2–20-passenger capacity and are better suited for destination resorts and smaller marine leisure locations. In 2005, an American company launched the *Alicia*, a 1000-ft (300-m) tourist submersible that can carry six passengers.

In addition to passenger-carrying tourist submarines, there are a number of personal or recreational submersible designs now being offered to the leisure market. Most of these are wet submersibles, although in 2005 two 1-atm DSVs were delivered to private owners.

Military. The most important applications are intelligence gathering, search, identification, location, retrieval, and neutralization of ordnances. These include explosive ordnance (mines, torpedoes, and bombs), downed aircraft, sunken vessels, and other objects of high national interest. Other military applications include hardware site and cable route survey, submarine rescue, wreck marking and destruction, topographic surveying, radiation measurements, water sampling, and water turbulence measurements and tests.

For several years, special warfare organizations, such as the U.S. Navy SEALS, have used small submersibles as swimmer delivery vehicles. In general, they are carried on submarines and the submersibles are launched and recovered while the submarine is submerged. In the larger navies, "wet" ambient-pressure vehicles are being replaced by 1-atm dry submersibles. However, many of the smaller navies in the world will continue to use the wet submersibles for covert operations.

Up through the 1970s, military research and development was the primary force behind the development of all types of submersibles. Today, almost all underwater submersible development is done in industry and academic research facilities. The military either contracts with them for services or buys this equipment from the private sector.

Scientific research. The crewed vehicle, particularly the 1-atm untethered variety, continues to be the workhorse of the scientific community. Twenty-four are in service worldwide for support of marine research. In this role, the submersible has become a very sophisticated platform from which the scientist can directly observe and measure a variety of environmental aspects, which are best understood by direct visual observation. The best-known is the Navy-owned *Alvin*, which has been operated by the Woods Hole Oceanographic Institution (WHOI) since 1964. Diving as deep as 13,000 ft (4000 m), *Alvin* has made many significant discoveries in the deep ocean. Since the early 1980s, a family of 20,000-ft (6000-m) DSVs have been used for research work. Examples of this class are the two Russian Mir submersibles in operation since 1987 (**Fig. 12**).

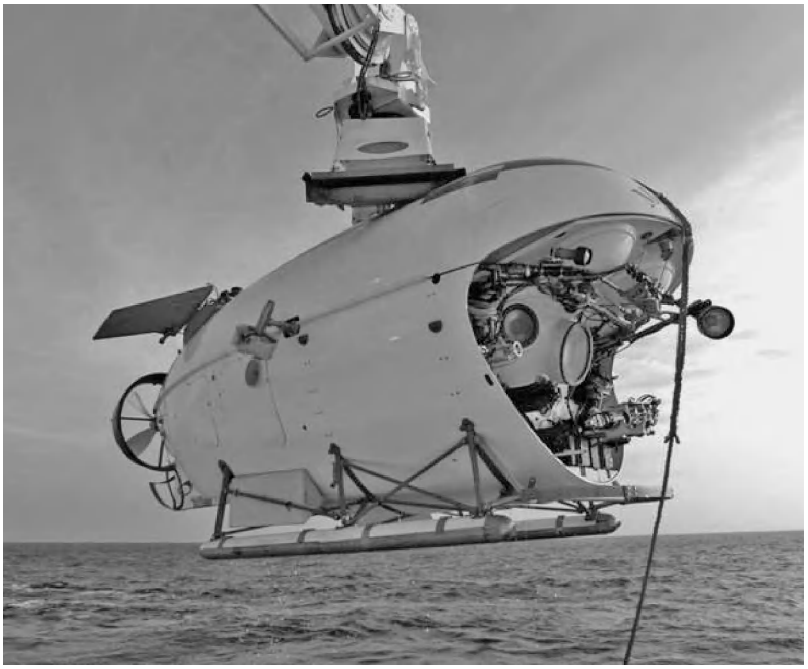


Fig. 12. One of the two Russian Mir 20,000-ft (6,000-m) manned submersibles. (*P.P. Shirshov Institute of Oceanology*)

In recent years, several scientific institutions have started to use remotely operated vehicles for research platforms. They offer an affordable submersible platform with excellent flexibility for a wide range of mission configurations. The rapid progress in development of microcomputers has made it possible to put much more capability into these small vehicles. *See* MICROPROCESSOR.

Towed vehicles have been used for research tasks since the 1960s. They are used to conduct deep-water geophysical and geological surveying and reconnaissance, midwater biological and water sampling, and collection of sea-floor sediments.

Some scientific organizations have combined the use of remotely operated and crewed vehicles to economize operations and at-sea time. For example, studies of hydrothermal vents by the crewed submersible *Alvin* were preceded by observing many square miles of sea floor with the towed vehicle *Angus* (which actually discovered the vents). *Alvin*, with its human observers, was subsequently deployed to conduct closeup observations and sampling, its primary role. Later, a small remotely operated vehicle, *Jason Jr.*, was used to swim off from *Alvin* to investigate places on the wreck that were too narrow or hazardous for the crewed vehicle to enter. This practice of using a towed vehicle

to perform a relatively high-speed, large-area reconnaissance and a crewed vehicle (sometimes with an on-board ROV) for final significant observations saves many days, perhaps months, of effort, since the crewed submersible is not an effective high-speed search and survey platform. *See* HYDROTHERMAL VENT.

The AUV has become a powerful new platform for oceanographic research where high-quality measurements are required but human presence is not. Eventually it should be possible for AUVs to make unattended transects of oceans, with the vehicle moving between a set depth and the surface while taking a variety of oceanographic measurements. Navigation updating and data transfer can be done by a radio link between the AUV and a satellite in space. As noted earlier, the major technical problem is getting enough power on board to support long-duration missions. In the 1990s, a Russian group proposed an AUV equipped with solar cells that can recharge batteries when the vehicle is on the surface transmitting data to a satellite. In 2003, a U.S. company put this type of AUV on the market.

In addition to operational AUVs, the future will bring crewed long-duration systems. Only one example exists at present, the U.S. Navy's *NR-1*. The small nuclear-powered *NR-1* can operate with mission times of up to a month.

U.S. Navy nuclear submarines have been used to do scientific research under the ice in the Arctic, the least-known ocean. However, the last of the fully ice-capable submarines was retired in 2000. Military submarines are not ideal for the long-duration scientific missions, while *NR-1* has relatively short endurance. The optimum platform would be a built-for-the-purpose nuclear submarine. In the early 1990s, a Canadian company and their Soviet Union partner proposed the development of *Ocean Shuttle*, a long-duration, multimission, crewed submarine analogous to NASA's space shuttle system. It was not built due to the end of the Cold War and a change in Canada's government. Such a submarine platform will be expensive to build and operate, but it offers the only means to do certain kinds of critical research in the world ocean. This type of underwater platform will eventually be developed because there are ocean work tasks that can only be done in this way.

Other uses. As the work capabilities of underwater vehicles have been proved, additional new opportunities have opened up. This is especially true in the uses of tethered, free-swimming ROVs. They now have a wide variety of civil uses such as fisheries resources management; inspection and maintenance of ships; law enforcement and public safety; marine pollution detection, remediation, and monitoring; treasure hunting; and underwater archeology. Many of these uses take advantage of vigorous price competition at the low end of the inspection vehicle market, where an ROV can be obtained for about $15,000. The evolution of the AUV will also follow a similar path to provide affordable systems to a wide community. While these economies are being realized, the top-end submersibles will increase in complexity, capability, and price. *See* SUBMARINE.

Don Walsh

Bibliography. American Bureau of Shipping, *Rules for Building and Classing Underwater Vehicles, Systems and Hyperbaric Facilities*, ABS, 2002; R. D. Ballard, *The Eternal Darkness: A Personal History of Deep-Sea Exploration*, Princeton University Press, 2000; W. J. Broad, *The Universe Below*, Simon & Schuster, New York, 1997; W. Forman, *The History of American Deep Submersible Operations*, Best Publishing, 1999; G. Griffiths (ed.), *Technology and Applications of Autonomous Underwater Vehicles*, CRC Press, 2002; *Jane's Underwater Technology 2005-06*, Janes Publishing, New York, 2005; Marine Board, National Research Council, *Underwater Vehicles and the National Needs*, National Academy Press, Washington, DC, 1996; G. Roberts et al. (eds.), *Guidance and Control of Underwater Vehicles 2003*, Elsevier, 2003; Underwater News & Technology, *Remote Operated Vehicles of the World 1998/99 Edition*, Technology Systems Corp., Stuart, FL, 1999.

Uninterruptible power supply

A source of high-quality alternating power to sensitive electrical equipment. An ideal uninterruptible power supply (UPS) provides a critical load with power when the main supply suffers from an over- or undervoltage condition causing interruption of the supply. It thus ensures power without break for the critical load. The UPS should be reliable and highly efficient, requiring low maintenance, low cost, and light weight. Uninterruptible power supplies were originally designed to furnish reliable and high-quality continuous power to large computer systems. They are now applied to a wide variety of critical equipment, such as medical facilities, life-support systems, financial transaction data storage and computer systems, telecommunications, industrial processing, and on-line management systems. UPS ratings range from a few hundred volt-ampere (VA) up to several thousand kilovolt-ampere (kVA). Three types of UPS systems are available: static, rotary, and hybrid static/rotary. *See* CIRCUIT (ELECTRICITY); ELECTRIC POWER TRANSMISSION.

Static UPS. Static UPS are available in ratings from 100 VA to 1 MVA, powering equipment ranging from personal computers and telecommunication systems, to medium power medical systems, to high power main systems. They are the most commonly used UPS systems, since they have high efficiency, high reliability, and low distortion. The disadvantages of static UPS are their poor performance with nonlinear and unbalanced loads, and high costs of achieving very high reliability. The principal configurations of static UPS systems are line-preferred, inverter-preferred, and line-interactive.

Line-preferred UPS. Line-preferred UPS are also referred to as standby or off-line UPS. This type of UPS is extensively employed to support different electrical loads. A line-preferred UPS is shown in **Fig. 1**. The system contains an ac/dc converter, a battery bank, a dc/ac converter, and static and transfer switches. There are three operating modes: normal mode, recovery mode, and backup mode.

In the normal mode of operation, power available on the ac line is normal, and the transfer and the static switches transfer power from the ac line to the load. The dc/ac converter and the load are in parallel. The dc/ac converter can be employed as an active filter by adding a simple control unit to the system, in the normal mode of operation. Hence, line-preferred UPS decreases the line current harmonics and improves the load power factor. Furthermore, the ac/dc converter can be controlled to reduce the harmonic content of the current on the dc side.

During recovery mode, the ac line directly feeds the load as well as the ac/dc converter to charge the battery. This mode can be activated with proper control strategy whenever needed. The ac/dc converter is not required to feed the load, because it is used only as a charger. Consequently, it allows low power rating, which makes it possible for the line-preferred UPS systems to be cheaper than the other static UPS topologies.

In backup mode, the ac line is not available, and the transfer switches disconnect the line from the ac/dc converter to prevent accidents. The dc/ac converter is turned on at the same time, and the battery unit supplies critical load via the converter until the main is within the preset tolerance again. The most important issue is the transfer time from normal mode to backup mode of operation. It is usually about $^1/_4$ line cycle, which is enough for most of the applications such as personal computers. *See* ALTERNATING CURRENT; CONVERTER; DIRECT CURRENT.

Inverter-preferred UPS. The inverter-preferred configuration, also called a double-conversion UPS or on-line UPS, is shown in **Fig. 2**. The system consists of an ac/dc converter (rectifier/charger), a battery unit, a dc/ac converter (inverter), and static (bypass) and transfer switches. The inverter-preferred UPS system has four operating modes: normal, recovery, stored energy, and bypass.

In the normal mode the transfer switch is on and the bypass switch is in the off position. The transfer switch continuously transfers power to the load from the dc/ac converter through the ac/dc converter. In this mode of operation, in contrast with the line-preferred UPS topology, the dc/ac and ac/dc converters are always on. Because double conversion occurs in this configuration, line conditioning is very good. The ac/dc converter is employed as a rectifier and charger for the load and battery. Hence, the ac/dc converter must be rated at 100% of the power demand of the load and battery unit. The dc/ac converter also must be at 100% of power load demand, since it is always on. Whenever the ac line is out of the tolerable limits, to avoid any undesirable power event between the ac line and the critical load, the transfer switch is off. This mode is

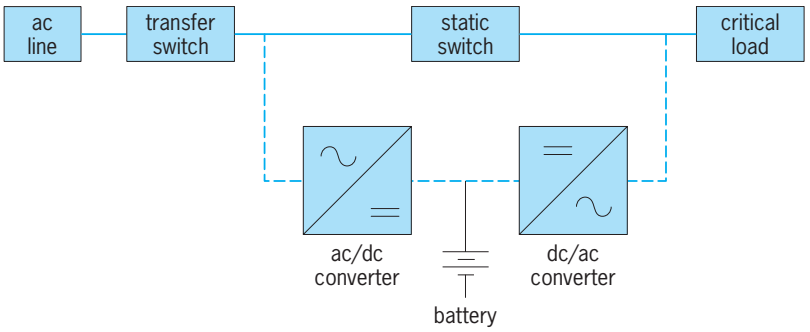


Fig. 1. Block diagram of a line-preferred UPS (off-line technology).

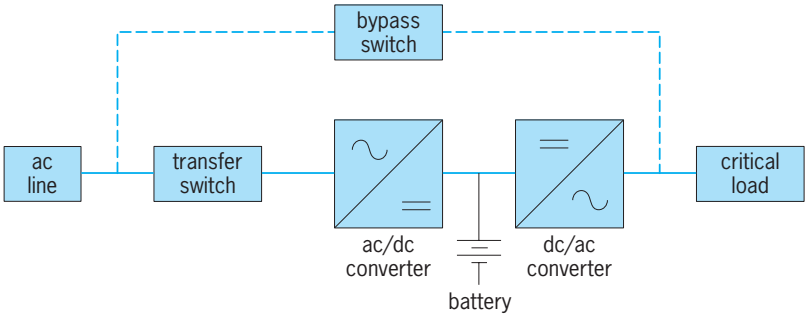


Fig. 2. Block diagram of an on-line UPS with bypass switch.

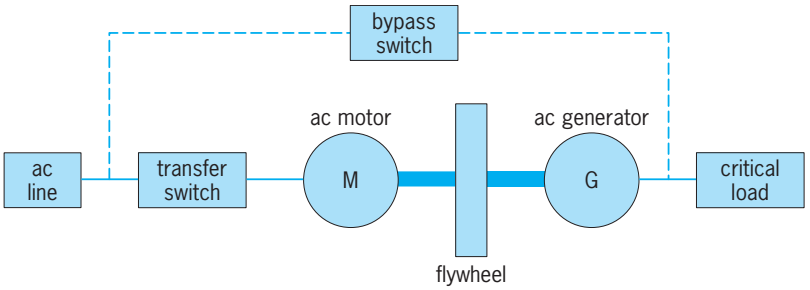


Fig. 3. Block diagram of the flywheel rotary UPS topology.

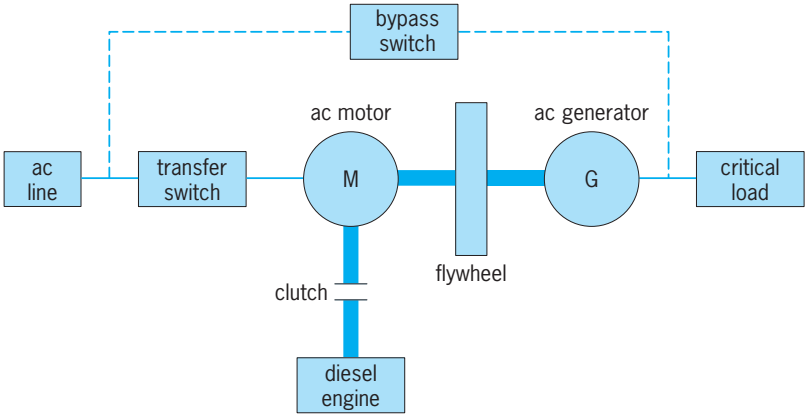


Fig. 4. Block diagram of diesel rotary UPS topology.

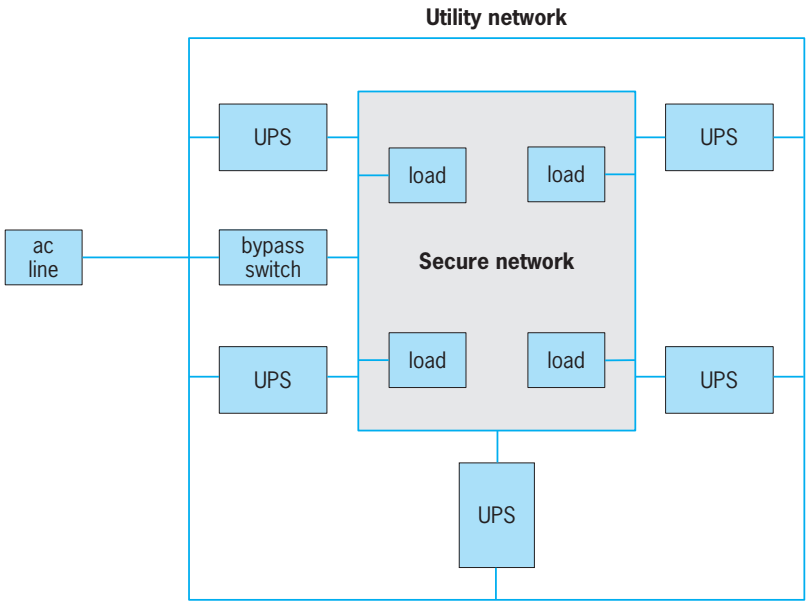


Fig. 5. Block diagram of the online distributed configuration.

known as backup mode. Since the dc/ac converter is always on, it continues to deliver power from the battery to the critical load without interruption. The UPS system returns to normal mode when the ac input voltage is within the preset tolerance again.

The main advantage of this configuration is that there is no transition time between normal and backup modes. Moreover, the inverter-preferred UPS system is more reliable than the line-preferred UPS topology; and, since the ac/dc converter is always on, the output frequency can be widely adjusted. Additionally, very wide tolerance to the input voltage differences and very good regulation of the output voltage can be realized by double-conversion topology.

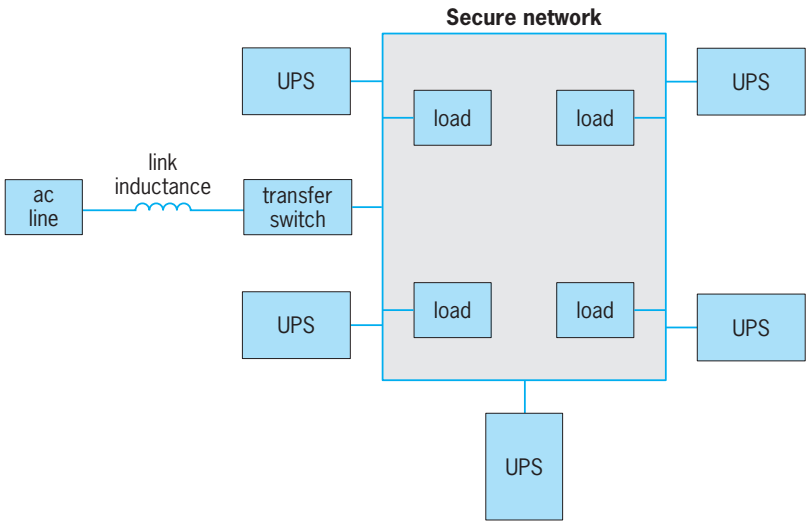


Fig. 6. Block diagram of line-interactive distributed application.

Line-interactive UPS. The Line-interactive UPS is also known as the grid-interactive UPS system. There are three types: single-conversion, in-line, and double-conversion topologies. They offer advantages such high efficiency and improved performance. For example, the single-conversion line-interactive UPS does not produce additional harmonics at the input side, a major advantage. High efficiency, low cost, and applicability in high power rating are the main advantages of the delta-conversion UPS configuration compared to the on-line UPS.

Rotary UPS. In rotary UPS systems a motor/generator set is used for energy conversion. There are two types of rotary UPS systems: flywheel rotary and diesel rotary. As shown in **Fig. 3**, a flywheel rotary UPS contains an ac motor, an ac generator, a flywheel unit, a static switch, and an optional bypass switch. The flywheel unit provides a mechanical coupling between the ac motor and ac generator, and is the energy storage component in this topology, rather than a battery as in the static UPS system. It provides dynamic energy for the generator and feeds the critical load through the generator when the ac line is out of preset tolerance. Restricted frequency range, restricted speed range, limitation on energy extraction and huge flywheel dimensions, even for a few seconds backup for the load, are the main disadvantages of this configuration.

The diesel rotary UPS is depicted in **Fig. 4**. The motor/generator set is connected to a diesel engine by means of a clutch. In normal mode, the ac line provides the power for the motor; the generator merely distributes the energy to the load. In backup mode, the ac motor is disconnected from the power line by the control unit, and the diesel engine begins to operate. It does not power the load at once, since it needs a certain amount of start-up time. The rotational energy stored in the flywheel unit provides interim power for the load until the diesel engine reaches its rated speed. After that, the clutch is engaged and the system operates as a diesel/generator set to supply the critical load. The flywheel unit does present disadvantages due to its diameter and weight. Nevertheless, the rotary UPS systems are definitely more reliable than static UPS systems. *See* ALTERNATING-CURRENT GENERATOR; CLUTCH; DIESEL ENGINE; FLYWHEEL; MOTOR.

Hybrid Static/Rotary UPS. Hybrid static/rotary UPS systems combine the features of both the static and rotary UPS systems. The main advantages of this configuration are better power quality with nonlinear loads, high reliability, frequency stability, better isolation, and low cost of maintenance are.

Applications. UPS systems provide uninterruptible and reliable power for the load including power conditioning in the system. There are two approaches to deploying UPS systems: distributed and centralized.

Distributed approach. Many parallel UPS units placed in an interconnected secure network feed the critical loads in a distributed system. Power circulates between the UPS units and the critical load

elastically. On-line and line-interactive approaches are two distributed schemes commonly used.

In the on-line distributed configuration, the secure and utility networks are connected to each other with the parallel UPS units, as shown in **Fig. 5**. There is isolation between the secure and utility networks which is provided by dc links of the UPS units. When the utility network is live, the critical loads are supplied by the ac line via the UPS units. The reactive current required by the secure network load is also provided by the UPS units. The utility network current harmonics are isolated from the secure network as well. Hence, the UPS units provide power conditioning and isolation for the system.

When the ac line is out of preset tolerance, the secure network is fed by the UPS batteries without any disruption. If one of the UPS units fails, in order to avoid power trouble, this unit is disconnected from the network instantly. Since power sharing is available between the parallel units, the other UPS units maintain power for the network, instead of failed UPS unit.

In the line-interactive distributed configuration, there is only one network instead of the utility and the secure networks, as shown in **Fig. 6**. Therefore, the system has a simpler structure compared to the online distributed topology, and the system cost is less, because the UPS units only work when the utility is in failure. Moreover, there is only one inverter in each UPS units. This characterization provides bidirectional power flow for the system. As mentioned before, there is no rectifier stage in the line-interactive distributed configuration; therefore the system is more reliable and more efficient than the online distributed approach.

When the utility is normal, the secure network is supplied by the ac mains. Simultaneously, the batteries are charged by the inverters. The reactive power demanded the secure network loads is also provided by the inverters. This configuration also achieves power line conditioning. Thanks to link inductance, the ac line absorbs all load changes with a delaying process (approximately 0.5 ms). The batteries feed the secure network loads in the backup mode of operation. The system has to provide seamless transition from the normal mode to the backup mode. To achieve this, the power angle is rapidly changed by the inverter from a negative value to a positive value. The inverter has a high switching frequency (10 kHz).

There are load sharing and system monitoring problems in this topology. Therefore, the system needs very fast digital controllers, which means high cost, a disadvantage of distributed UPS systems.

Centralized approach. Only one large UPS unit is employed in the centralized configuration, as shown in **Fig. 7**. The large UPS provides continuous operation for the whole system. This configuration is more desirable for industrial and utility applications, due to low maintenance requirements. To overcome the redundancy problem and to expand load capacity, additional costs are incurred in the centralized approach. Moreover, service and maintenance specialists are needed, also contributing to increase costs. However, such specially trained groups decrease the risk and increases the reliability of the system.

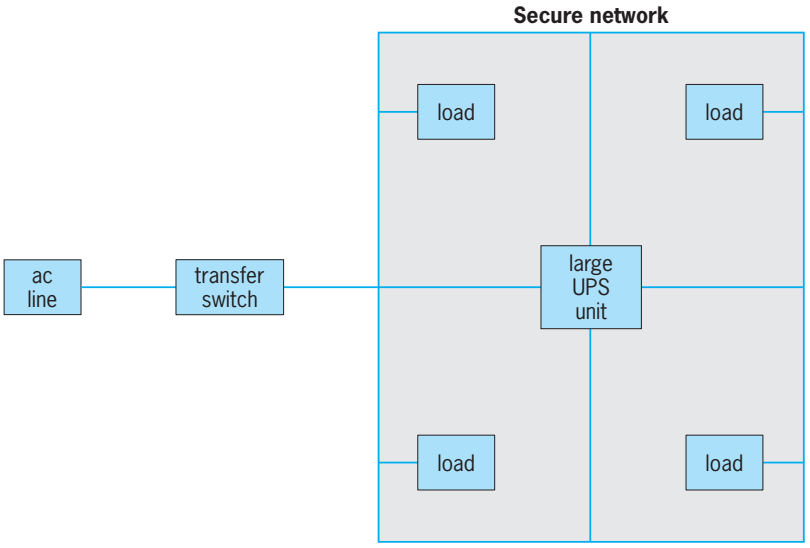


Fig. 7. Block diagram of the centralized approach.

Ayse E. Amac; Ali Emadi

Bibliography. S. B. Bekiarov and A. Emadi, Uninterruptible power supplies: Classification, operation, dynamics, and control, in *Proceedings of the 17th Annual IEEE Applied Power Electronics Conference and Exposition*, Dallas, TX, March 2002, pp. 597–604; M. C. Chandorkar, Control of distributed UPS systems, in *Proceedings of the 25th Annual IEEE Power Electronics Specialists Conference*, Taipei, Taiwan, June 1994, pp. 197–204; S. A. O. da Silva et al., A three phase line-interactive UPS system implementation with series-parallel active power line conditioning capabilities, *IEEE Trans. Ind. Applicat.*, 38(6):1581–1590, 2002; W. W. Hung and G. W. A. McDowell, Hybrid UPS for standby power systems, *Power Eng. J.*, 4(6):281–291, 1990; A. A. McLennan, Static UPS technologies, *IEE Colloquium on Uninterruptible Power Supplies*, London, February 1994, pp. 2/1–2/5; S. R. Philpott, Large scale UPS installations, in *Proceedings of the IEEE International Conference on Electrical Installation Engineering in Europe*, London, June 1993, pp. 64–68; J.-C. Wu and H.-L. Jou, A new UPS scheme provides harmonic suppression and input power factor correction, *IEEE Trans. Ind. Elec.*, 42(6):629–635, 1995.

Unit operations

A structure of logic used for synthesizing and analyzing processing schemes in the chemical and allied industries, in which the basic underlying concept is that all processing schemes can be composed from and decomposed into a series of individual, or unit, steps. If a step involves a chemical change, it is called a unit process; if physical change, a unit operation.

These unit operations are utilized for certain definite functions wherever they are employed, and their use cuts across widely different processing applications, including industries as diverse as manufacture of chemicals, fuels, pharmaceuticals, pulp and paper, processed foods, and primary metals. The unit operations approach thereby serves as a very powerful form of morphological analysis, which systematizes process design, and greatly reduces both the number of concepts that must be taught and the number of possibilities that should be considered in synthesizing a particular process.

The selection and classification of unit operations is somewhat arbitrary. Filtration is generally thought of as a single unit operation, but in reality there are a number of quite different types of filters. Distillation, solvent extraction, and adsorption are usually regarded as different unit operations, but all can serve to fulfill the same function of separating a liquid mixture into two or more products of different composition. Distillation accomplishes this through differences in volatility, solvent extraction utilizes different solubilities into an immiscible solvent, and adsorption is based on different affinities for a solid surface. All fall within the general category of separation processes, which is itself a focused area of investigation. Different separation techniques are useful or preferred for different mixtures and purposes. The invention and selection of separation techniques is an important dimension of process synthesis, along with selection among, and assemblage of, other unit operations.

Most of the unit operations are based mechanistically upon the fundamental transport processes of mass transfer, heat transfer, and fluid flow (momentum transfer). Unit operations based on fluid mechanics include fluid transport (such as pumping), mixing/agitation, filtration, clarification, thickening or sedimentation, classification, and centrifugation. Operations based on heat transfer include heat exchange, condensation, evaporation, furnaces or kilns, drying, cooling towers, and freezing or thawing. Operations that are based on mass transfer include distillation, solvent extraction, leaching, absorption or desorption, adsorption, ion exchange, humidification or dehumidification, gaseous diffusion, crystallization, and thermal diffusion. Operations that are based on mechanical principles include screening, solids handling, size reduction, flotation, magnetic separation, and electrostatic precipitation. The study of transport phenomena provides a unifying and powerful basis for an understanding of the different unit operations. *See* CHEMICAL ENGINEERING; TRANSPORT PROCESSES; UNIT PROCESSES.

For further information concerning operations based on fluid mechanics *see* CENTRIFUGATION; CLARIFICATION; FILTRATION; MIXING; PUMP; PUMPING MACHINERY; SEDIMENTATION (INDUSTRY); THICKENING. For operations based on heat transfer *see* COOLING TOWER; DRYING; EVAPORATION; HEAT EXCHANGER; KILN. For operations based on mass transfer *see* ABSORPTION; ADSORPTION; CRYSTALLIZATION; DEHUMIDIFIER; DIFFUSION; DISTILLATION; HUMIDIFICATION; ION EXCHANGE; LEACHING; SOLVENT EXTRACTION. For operations based on mechanical principle *see* ELECTROSTATIC PRECIPITATOR; FLOTATION; MAGNETIC SEPARATION METHODS; MECHANICAL SEPARATION TECHNIQUES; SOLIDS PUMP.

C. Judson King

Bibliography. W. F. Furter (ed.), *A Century of Chemical Engineering*, 1982; A. Grandison and M. Lewis, *Separation Processes*, 1995; D. Hagerty, E. Gearhard, and C. Plank, *Chemical Engineering*, 1989; C. J. King, *Separation Processes*, 2d ed., 1980; W. L. McCabe and J. C. Smith, *Unit Operations of Chemical Engineering*, 5th ed., 1993; R. Reynolds, *Unit Operations and Processes in Environmental Engineering*, 2d ed., 1996.

Unit processes

Processes that involve making chemical changes to materials, as a result of chemical reaction taking place. For instance, in the combustion of coal, the entering and leaving materials differ from each other chemically: coal and air enter, and flue gases and residues leave the combustion chamber. Combustion is therefore a unit process. Unit processes are also referred to as chemical conversions.

Together with unit operations (physical conversions), unit processes (chemical conversions) form the basic building blocks of a chemical manufacturing process. For instance, reactants are mixed together (unit operation) and allowed to react (unit process), and the product stream is separated into its constituents which then may or may not be purified further (various unit operations). Most chemical processes therefore consist of a combination of various unit operations and unit processes.

The basic tools of the chemical engineer for the design, study, or improvement of a unit process are the mass balance, the energy balance, kinetic rate of reaction, and position of equilibrium (the last is included only if the reaction does not go to completion). Together these are popularly called the four horsemen of chemical engineering, and the study or design is not complete unless all four have been addressed.

A list of unit processes can be neither rigid nor permanent. New ones are added as they are developed. A partial list follows (see separate articles on the items with asterisks):

Acetylation*
Acidolysis
Acidulation
Acylation*
Addition reaction*
Alcoholysis*
Alkali fusion reaction*
Alkylation*
Amination*
Ammoniation
Ammonolysis
Aromatization*
Calcination
Carboxylation
Causticization
Chelation*
Combustion*
Condensation reaction*
Consolidation reaction
Corrosion
Cyanoethylation*
Cyclization

Dealkylation
Decomposition
Dehydration reaction
Dehydrogenation*
Diazotization* and coupling
Double decomposition
Electrolysis*
Epoxidation*
Esterification
Etherification
Fermentation*
Fischer-Tropsch process*
Friedel-Crafts reaction*
Grignard reaction*
Haloform reaction*
Halogenation*
  bromination
  chlorination
  fluorination
  iodination
Halogenolysis
  brominolysis
  chlorinloysis
  fluorinolysis
  iodinolysis*
Hydration*
Hydroboration*
Hydrodealkylation
Hydroformylation*
Hydrogenation*
Hydrogenolysis
Hydrolysis*
Hydrorefining
  hydrodesulfurization
  hydrofining
Hydroxylation reaction*
Interesterification
Ion exchange*
Isomerization*
Methanation
  methanization
Neutralization
Nitration
Nitrogen fixation*
Oxidation process*
  also bleaching*
  detonation
  explosion and explosive*
Oxoformylation
Ozonolysis*
Photolysis*
Photosynthesis*
Polymerization*
  addition polymerization
  condensation polymerization
  dimerization
  photopolymerization
  trimerization
Polyoxyethylation
Pyrolysis*
  carbonization
  coking
  cracking
  destructive distillation
  dry distillation
Racemization*
Reduction
Reforming processes*
  hydroforming
  platforming
Saccharification
Saponification
Silicate formation
Substitution reaction*
Sulfonation*
Transesterification*
Vulcanization

See CHEMICAL ENGINEERING; UNIT OPERATIONS.

W. F. Furter

Unitary symmetry

A type of symmetry law, an important example of which is flavor symmetry, an approximate internal symmetry law obeyed by the strong interactions of elementary particles. According to the successful theory of strong interactions, quantum chromodynamics, flavor symmetry is the consequence of the fact that the strong force (the so-called glue force, mediated by the SU_3^{color} gauge field) is the same between any two quarks, the constituents of hadrons. It follows that if all the kinds (flavors) of quarks had the same mass, strong interactions would have the symmetry SU_N^{flavor}, where N is the number of quark flavors. As a result, if nonstrong (electromagnetic and weak) interactions were neglected, then hadrons would occur as degenerate (all having the same mass) multiplets (irreducible representations) of the group SU_N. *See* COLOR (QUANTUM MECHANICS); FLAVOR; HADRON; QUANTUM CHROMODYNAMICS; QUARKS.

Spin independence. An example of unitary symmetry is the spin independence of electric forces on slowly moving electrons. The electron has spin $^1/_2$ and thus two spin states, referred to as spin-up and spin-down. Denoting the wave functions of these two states by $|u\rangle$ and $|d\rangle$, the spin-independent physical properties (energy eigenvalues, charge density, and so on) of a system of electrons such as those in an atom, in which the forces on electrons are nearly spin independent, are unchanged by the replacements shown in Eqs. (1), where α and β are

$$\begin{aligned} |u\rangle &\rightarrow \alpha|u\rangle + \beta|d\rangle \\ |d\rangle &\rightarrow -\beta^*|u\rangle + \alpha^*|d\rangle \\ |\alpha|^2 + |\beta|^2 &= 1 \end{aligned} \tag{1}$$

complex numbers. These transformations form a group known as SU_2. They correspond to rotations of space; α and β can be expressed in terms of the three numbers which describe the rotation. It is easily seen that many-electron states decompose under the transformation; for example, using Eqs. (1), the two-electron state $(|u,d\rangle - |d,u\rangle)/\sqrt{2}$ is unchanged, and the three remaining two-electron states, $|u,u\rangle$, $|d,d\rangle$, and $(|u,d\rangle + |d,u\rangle)/\sqrt{2}$, transform to linear combinations of themselves. This is the decomposition into singlet and triplet spin states, that is, into total spin $S = 0$ and 1, respectively; the nonmixing between them is equivalent to the invariance of S to rotation. *See* ANGULAR MOMENTUM; ELECTRON SPIN; SPIN (QUANTUM MECHANICS).

SU_2 flavor symmetry (charge independence). The first flavor symmetry encountered in particle physics was the near equivalence of the proton and neutron in nuclear physics. This is called charge independence, *i*-spin (isotopic spin) invariance, or (in terms of the quark model) flavor SU_2 symmetry. The nucleons, proton and neutron, both have spin $^1/_2$ and have nearly the same mass (differing by only 0.14%), and the nuclear force between any two nucleons—whether two protons, a proton and a neutron, or two neutrons—is nearly the same. This is just like the approximate spin independence of the force between two electrons and is describable in the same way, as an SU_2 symmetry, Eqs. (1). Hence one speaks of the proton and neutron as being the two *i*-spin states of a nucleon, a particle with *i*-spin $^1/_2$. Conventionally, the proton is the *i*-spin up state and the neutron is *i*-spin down. For real spin in empty space, spin-up and spin-down are defined with respect to an arbitrary direction [if the direction is changed, the spin-up and spin-down states change by the transformation Eqs. (1)], but for *i*-spin electromagnetism permanently singles out a direction in "*i*-spin space," the nucleon's *i*-spin up state (proton) having charge e and its *i*-spin down state (neutron) having charge zero. A real-spin analog of this would be for a spin-$^1/_2$

particle carrying a magnetic moment to be always in a constant magnetic field.

The states of nuclei (bound systems of nucleons) can be approximately described as having a definite total *i*-spin (*I*), but since the electrostatic energy of a nucleus rises with the square of the number of its protons (*Z*) this description rapidly becomes worse for higher-*Z* nuclei.

The *i*-spin symmetry extends to all the strongly interacting particles, both baryons and mesons, all of which have *i*-spin as one of their properties. An unsuccessful attempt was made to explain this universality by supposing, for example, that the pion (pi meson), which has *i*-spin 1, was a nucleon-antinucleon bound state. But the universality is now explained in the quark model of hadrons by the fact that the two least massive quarks, the up (*u*) and down (*d*) quarks, are similar (analogous to the similarity of proton and neutron). The masses of the *u* and *d* quarks are estimated at roughly 3 MeV and 6 MeV, respectively (the neutron is heavier than the proton because the *d* is heavier than the *u*). Despite this mass ratio of 2 the *u* and *d* really do behave similarly in a hadron, because they have there a zero-point energy of the order of 200 or 300 MeV, and so the effect of their masses on their dynamics is very small.

The *i*-spin of any hadron is thus determined by its *u* and *d* quarks. For instance, the lightest (*s*-wave) $J^P = {}^3/_2{}^+$ baryons contain three quarks in a symmetric flavor state (this is required to make the overall state of the three fermions symmetric under exchange), and therefore they have *i*-spin ${}^3/_2$, 1, ${}^1/_2$, 0 according to whether 3, 2, 1, 0 of their quarks are *u* or *d*. Similarly, exchange symmetry implies that the lightest $J^P = {}^1/_2{}^+$ baryons have *i*-spin ${}^1/_2$, 0 or 1, or ${}^1/_2$ according to whether 3, 2, or 1 of their quarks are *u* or *d*. *See* I-SPIN.

SU_3 flavor symmetry. The third lightest quark is the strange quark (*s*), with mass about 0.15 GeV. It is named in reference to the "strange particles" (hyperons and *K* mesons), which were found to be produced in high-energy collisions almost as copiously as pions but did not decay rapidly into nucleons and pions despite being massive enough to do so. In terms of the quark model, one of the quarks (or antiquarks) of a strange particle is an *s* quark (or anti-*s* quark). (There are also doubly and triply strange baryons, in which two or all three of the quarks are *s* quarks.) In the decay of a strange particle into nonstrange particles (or of a doubly strange particle into a singly strange particle, and so on) an *s* quark must change into a *u* or *d*; this can happen only (as with any flavor change of a quark) by a weak interaction, not a strong interaction, and so the decay is slow. This is similar to ordinary beta decay of a neutron, which involves the weak-interaction change of a *d* quark into a *u*. *See* WEAK NUCLEAR INTERACTIONS.

If the *s* quark were as light as a *u* or *d* quark, then these three kinds would be equivalent (as far as the strong interactions were concerned), and any flavor-independent physical properties (energy eigenvalue, for instance) of a state of quarks would be unchanged by replacements similar to the SU_2 transformation shown in Eqs. (1) but with a third state, $|s\rangle$, involved in addition to $|u\rangle$ and $|d\rangle$, and with the complex coefficients specified by eight (rather than three) real numbers. These symmetry transformations form a group known as SU_3 [SU_N denotes the group of $N \times N$ unimodular (determinant = 1) unitary matrices]. The triplet of states $|u\rangle$, $|d\rangle$, and $|s\rangle$, equivalent under the SU_3 symmetry, is analogous to the doublet $|u\rangle$ and $|d\rangle$, equivalent under the SU_2 (*i*-spin) symmetry. [It is not to be confused with an *i*-spin triplet of SU_2. The group of rotations of space is essentially the same group as the group SU_2, which is the reason for the exact analogy described above between a real-spin doublet and a flavor SU_2 doublet (both denoted by $|u\rangle$ and $|d\rangle$), and for the use of the term *i*-spin. SU_3 is different.] The baryons described at the end of the previous section provide examples of SU_3 multiplets if the quarks in the baryon are *u*, *d*, or *s*, namely the octet and the decuplet (**Table 1**). *See* STRANGE PARTICLES.

TABLE 1. Lightest (*s*-wave) baryons made of *u*, *d*, or *s* quarks

Quark content*	$J^P = {}^1/_2{}^+$		$J^P = {}^3/_2{}^+$	
	I-spin	Names†	*I*-spin	Names
nnn	${}^1/_2$	$N^{0,+}$	${}^3/_2$	$\Delta^{-,0,+,++}$
nns	0; 1	Λ^0; $\Sigma^{-,0,+}$	1	$\Sigma^{-,0,+}$
nss	${}^1/_2$	$\Xi^{-,0}$	${}^1/_2$	$\Xi^{-,0}$
sss			0	Ω^-
	$[SU_3(8)]$		$[SU_3(10)]$	

*n denotes a nucleon-type quark, that is, *u* or *d*.
†N^0, N^+ are neutron, proton.

Because the *s* quark is significantly heavier than the *u* and *d* quarks, the SU_3 (*u, d, s*) symmetry is only approximate; the masses of the hadrons belonging to an SU_3^{flavor} multiplet differ from one another by the order of 100 or 200 MeV. Sometimes, as in the lightest $J^p = 1^-$ meson and ${}^3/_2{}^+$ baryon multiplets, the masses depend nearly linearly on the number of *s* quarks contained, as one would naively expect. In other cases, such as the lightest ${}^1/_2{}^+$ baryon multiplet, the mass of the *s* also has an indirect effect on the mass of the baryon through the strength of the spin-spin color interaction (the color analog of the hyperfine interaction, that is, the interaction of magnetic moments) because the color-magnetic moment of the heavier *s* is smaller than that of the *u* or *d*, just as for ordinary magnetic moments; this is also the explanation of the difference of the Λ and Σ masses.

The SU_3^{flavor} symmetry implies relations between strong coupling constants (for example, $\pm\sqrt{3}g_{KN\Lambda} \pm g_{KN\Sigma} = 2g_{\pi NN}$) and between cross sections; those which are testable experimentally seem to agree. There are also many relations among the matrix elements of electromagnetic and weak current operators. For instance, the magnetic moments of the ${}^1/_2{}^+$ baryon octet are predicted to have the relations given by Eqs. (2), according to which all moments of the octet members follow from the

TABLE 2. The lightest (*s*-wave) baryons made of *u*, *d*, *s*, or *c* quarks

Quark content*	$J^P = {}^1/_2{}^+$		$J^P = {}^3/_2{}^+$	
	SU_3 multiplet†	Names	SU_3 multiplet†	Names
lll	(8)	N, Λ, Σ, Ξ	(10)	$\Delta, \Sigma, \Xi, \Omega$
llc	$(\bar{3})$; (6)	$\Lambda_c, \Xi_c; \Sigma_c, \Xi'_c, \Omega_c$	(6)	$\Sigma_c, \Xi_c, \Omega_c$
lcc	(3)	Ξ_{cc}, Ω_{cc}	(3)	Ξ_{cc}, Ω_{cc}
ccc			(1)	Ω_{ccc}
		$[SU_4(20')]$		$[SU_4(20)]$

*l denotes a light quark, that is, u, d, or s.
†The SU_3 multiplets (1), (3), $(\bar{3})$, (6), (8), and (10) are also called singlet, triplet, antitriplet, sextet, octet, and decuplet, respectively.

proton and neutron moments.

$$\begin{aligned} \mu(\Sigma^+) &= \mu(p) \\ 2\mu(\Lambda) = \mu(\Xi^0) &= -2\mu(\Sigma^0) = \mu(n) \\ \mu(\Sigma^-) = \mu(\Xi^-) &= -[\mu(p) + \mu(n)] \end{aligned} \qquad (2)$$

These relations are found to be qualitatively correct, although the effect of the smaller magnetic moment of the *s* quark is clearly seen: For example, twice the ratio of the moments of the Λ and the neutron is 0.64; according to the naïve quark model, this is the ratio of the moments of the *s* and *d* quarks. *See* BARYON; MESON; WEAK NUCLEAR INTERACTIONS.

Just as the SU_2 (*u, d* or *i*-spin) symmetry applies to all hadrons, so does the (more approximate) SU_3 (*u, d, s*) symmetry. For instance, the lightest (*s*-wave) baryons whose quarks are *u*, *d*, *s*, or *c* (the next heavier quark) are described in **Table 2**. [In the naming scheme of these baryons, the capital letters specify the *i*-spin as in the $SU_3(8)$ and $SU_3(10)$ multiplets of Table 1. The *c* subscripts specify how many *s* quarks have been replaced by *c* quarks; each such replacement raises charge by *e* (the charge difference of *c* and *s*). For example, whereas the members of a Ξ multiplet are $\Xi^{-,0}$, those of a Ξ_{cc} are $\Xi_{cc}{}^{+,++}$. The two ${}^1/_2{}^+$ baryons with identical quantum numbers, Ξ_c and Ξ'_c, have different spin-flavor structure and differ in mass (by 0.11 GeV) because the color-magnetic moments of the *n*, *s*, and *c* quarks differ from one another.] Of these lightest ${}^1/_2{}^+$ and ${}^3/_2{}^+$ baryons, only those with more than one *c* quark have not yet been seen.

$SU_{4,5,6}$ flavor symmetry. The lightest of the three so-called heavy quarks, the charmed quark (*c*), has a mass of about 1.25 GeV. This large mass badly breaks the SU_4 flavor symmetry that would hold if *u, d, s*, and *c* all had the same mass. Although the lightest ${}^1/_2{}^+$ and ${}^3/_2{}^+$ baryons listed in Table 2 can each be described as comprising a 20-member multiplet of SU_4, SU_4 symmetry is not a useful starting point to relate properties of the multiplet members. The remaining heavy quarks, the bottom (*b*) and top (*t*) quarks, are heavier still with masses of about 4.5 and 175 GeV, respectively, and so SU_5 and SU_6 flavor symmetries are even more pointless to consider. In fact the top quark is so massive that it decays (weakly) before it has a chance to be a constituent of any well-defined hadron, its mean life being less than 10^{-24} s. *See* CHARM; J/PSI PARTICLE; UPSILON PARTICLES.

Chiral U_N^L U_N^R symmetry. If quarks were massless, their chirality would be a good quantum number, conserved by the glue interaction, and the flavor symmetry group would be larger, namely $U_N^L \times U_N^R$, where U_N^L is the flavor symmetry group of the left-handed quarks and U_N^R correspondingly for the right-handed quarks. Not only is this symmetry broken by the quark masses, but there is evidence (partly from numerical lattice quantum chromodynamics calculations) that it is spontaneously broken. In the approximate model where the masses of the *u* and *d* quarks are taken to be zero and the heavier quarks are ignored, the consequences of the spontaneous symmetry breaking of the $U_2^L \times U_2^R$ symmetry down to U_2 were worked out, in the σ model and current algebra, even before the quark model was developed, with results such as the Goldberger-Treiman relation between the strong and weak coupling constants of the charged π meson and the Adler-Weisberger relation between the Gamow-Teller and Fermi weak coupling constants of the nucleon.

Other unitary symmetries. There are unitary symmetries that are not purely internal in the strict sense, namely those in which it is supposed that the interaction between the fundamental particles is spin-independent (as for nonrelativistic electrons interacting with electrostatic forces). This was suggested by E. Wigner as an approximate symmetry in light nuclei; there the fundamental particle is a quartet, with states $|p\uparrow\rangle, |p\downarrow\rangle, |n\uparrow\rangle, |n\downarrow\rangle$(where $|p\uparrow\rangle$ means a proton with spin up, and so on), and the symmetry group is therefore SU_4. Similarly, SU_6 symmetry, resulting from combining flavor SU_3 with spin independence, is sometimes used in the quark model. It is known that such unitary symmetries can never be exact in a relativistic theory, but that fact does not prevent them from being approximately true for low-lying states. *See* ELEMENTARY PARTICLE.

Charles J. Goebel

Bibliography. M. Gell-Mann and Y. Ne'eman (eds.), *The Eightfold Way*, rev. ed., Perseus Publishing, 2000; G. Kane, *Modern Elementary Particle Physics*, Westview Press, 1993; D. B. Lichtenberg, *Unitary Symmetry and Elementary Particles*, 2d ed., Academic Press, 1978; U. Mosel, *Fields, Symmetries, and Quarks*, 2d ed., Springer-Verlag, 1999; L. H. Ryder, *Elementary Particles and Symmetries*, rev. ed., Gordon and Breach, 1986; N. P. Samios, M. Goldberg, and B. T. Meadows, Hadrons and SU(3): A critical review, *Rev. Mod. Phys.*, 46:49–81, 1974.

Units of measurement

Values, quantities, or magnitudes in terms of which other such are expressed. Units are grouped into systems, suitable for use in the measurement of physical quantities and in the convenient statement of laws relating physical quantities. A quantity is a measurable attribute of phenomena or matter.

A given physical quantity A, such as length, time, or energy, is the product of a numerical value or measure $\{A\}$ and a unit $[A]$. Thus Eq. (1) holds.

$$A = \{A\}[A] \tag{1}$$

The unit $[A]$ can be chosen arbitrarily, but it is desirable to define units in such a way that they are derived from a few base units by equations without numerical factors other than unity, and that the equations between numerical values of quantities have exactly the same form as the equations between the quantities. For example, the kinetic energy E of a body is given in terms of its mass M and speed V by Eq. (2), where $E = \{E\}\ [E]$, $M = \{M\}\ [M]$, $V = \{V\}\ [V]$, and $^1/_2$ is called a definitional factor and is dimensionless. If the units of E, M, and V are defined in such a way that Eq. (3) holds, then the equation between the numerical values is Eq. (4). A system

$$E = {}^1/_2 MV^2 \tag{2}$$

$$[E] = [M][V]^2 \tag{3}$$

$$\{E\} = {}^1/_2\{M\}\{V\}^2 \tag{4}$$

of units defined in this way is called a coherent system. It is constructed by defining the units of a few base quantities independently; these are called base units. The units of all other quantities are defined by equations similar to Eq. (3) with no numerical factors other than unity, and are called derived units. It is a matter of choice which quantities are to be considered as the base quantities and also how many. It is desirable to have as few base quantities as possible, provided no serious inconveniences are encountered. *See* DIMENSIONAL ANALYSIS.

In 1960 the General Conference on Weights and Measures (CGPM) gave official status to a single practical system, the International System of Units, abbreviated SI in all languages. The system is a modernized version of the metric system. The SI, as subsequently extended, includes seven base units, two supplementary units, and nineteen derived units with special names. These derived units, and others without special names, are derived from the base and supplementary units in a coherent manner. A set of prefixes is used to form decimal multiples and submultiples of the SI units. *See* METRIC SYSTEM.

The CGPM and its subsidiary body, the International Committee for Weights and Measures (CIPM), have accepted certain units which are not part of the SI but which are widely used or are useful in specialized fields, for use with the SI or for temporary use in those fields. For definitions of the seven SI base units, discussion of the physical standards for realizing these units, and tables of the SI base, supplementary, and derived units and units accepted for use with the SI *see* PHYSICAL MEASUREMENT.

This article discusses how the various derived units are defined in terms of the base units, both in the SI and in other coherent systems, and how units in different systems, as well as arbitrary units which are not part of any system, are related. Various units which are not accepted for use with the SI, but whose use is nevertheless common, are discussed. However, the use of all units which are not SI units or specifically accepted for use with the SI has been discouraged, and whenever possible, units outside the SI should be replaced by SI units or by their multiples and submultiples formed by attaching SI prefixes.

Geometrical Units

Units of plane angle and solid angle are purely geometrical. The SI units of plane and solid angle are regarded as dimensionless derived units.

Plane angle units. The radian (rad), the SI unit of plane angle, is the plane angle between two radii of a circle which cut off on the circumference an arc equal in length to the radius. Since the circumference of a circle is 2π times the radius, the complete angle about a point is 2π rad. *See* RADIAN MEASURE.

The degree and its decimal submultiples can be used with the SI when the radian is not a convenient unit. By definition, 2π rad $= 360^\circ$. The minute $[1' = (1/60)^\circ]$ and the second $[1'' = (1/60)']$ can also be used.

Steradian. The steradian (sr), the SI unit of solid angle, is the solid angle which, having its vertex at the center of a sphere, cuts off an area on the surface of the sphere equal to that of a square with sides of length equal to the radius of the sphere. Since the total area of a sphere is 4π times the square of its radius, the complete solid angle about a point is 4π sr.

Mechanical Units

In mechanics, it is convenient to have three base quantities, and two of these are generally chosen to be length and time. Systems of mechanical units may be classified as absolute systems, in which the third base quantity is mass, and gravitational systems, in which the third base quantity is force.

Absolute systems. Two absolute systems of metric units are commonly employed, each named for its base units of length, mass, and time: the mks (meter-kilogram-second) absolute system, and the cgs (centimeter-gram-second) absolute system. The mks absolute system is the mechanical portion of the SI. Once the length, mass, and time units are selected, Newton's second law in the form $F = ma$ is used to define the unit of force F in terms of mass (m) units and acceleration (a) units, which involve length and time units. The unit of force derived in this way through Newton's second law is called the newton in the mks system and the dyne in the cgs system. Units of work W are defined by the relation

TABLE 1. Absolute systems of units

Quantity	Defining relation	Units		
		mks (SI)	cgs	British
Length (L)	Fundamental	m	cm	ft
Mass (m)	Fundamental	kg	g	lb
Time (t)	Fundamental	s	s	s
Velocity (v)	$v = L/t$	m/s	cm/s	ft/s
Acceleration (a)	$a = v/t$	m/s^2	cm/s^2	ft/s^2
Force (F)	$F = ma$	1 N = 1 kg · m/s^2	1 dyne = 1 g · cm/s^2	1 pdl = 1 lb · ft/s^2
Work (W)	$W = FL$	1 J = 1 N · m	1 erg = 1 dyne · cm	ft · pdl
Power (P)	$P = W/t$	1 W = 1 J/s	erg/s	ft · pdl/s
Momentum (p)	$p = mv$	kg · m/s	g · cm/s	lb · ft/s

$W = FL$, where L represents a displacement in the direction of a force F. Units of power P in turn are defined by the relation $P = W/t$ and are therefore the ratio of work units to units of time t. A coherent absolute system of British units is based on the foot, the pound (1 lb $\cong$ 0.4536 kg), and the second; and the derived unit of force is called the poundal (pdl), a unit which is not often used. Various quantities expressed in the three absolute systems are listed in **Table 1**. *See* KINETICS (CLASSICAL MECHANICS); NEWTON'S LAWS OF MOTION.

Gravitational systems. Gravitational systems, in which the base quantities are length, force, and time, have been frequently employed by engineers, and are therefore sometimes called technical systems.

In the British gravitational system the standard force, called the pound force (lbf), is the gravitational force exerted on a 1-lb mass at a location on the Earth's surface where the acceleration of a freely falling body due to gravity has the standard value $g_n = 9.80665\ m/s^2 \cong 32.174\ ft/s^2$. More simply stated, 1 lbf is the force which imparts an acceleration of magnitude g_n to a 1-lb mass. A coherent system is obtained by using Newton's second law, $F = ma$, to define a mass unit called the slug (1 lbf = 1 slug · ft/s^2). The work unit in this system is the foot-pound force (ft · lbf). *See* FREE FALL.

In the mks gravitational system the unit of acceleration is the meter per second squared (m/s^2), and the unit of force is taken to be the Earth's gravitational force on a 1-kg mass at a location where this force imparts an acceleration of magnitude g_n. Unfortunately, this unit of force has also been called a kilogram, but to distinguish it from the unit of mass it should be called kilogram force (kgf). There is no generally accepted name for the mks gravitational unit of mass, although the name metric slug has been used. The cgs gravitational system is constructed in a similar manner. Various quantities expressed in the three gravitational systems are listed in **Table 2**.

Noncoherent gravitational systems in which force is expressed in pounds force (lbf) or a comparable kilogram force (kgf) or gram force (gf), and mass in pounds, kilograms, or grams, can be set up by writing Newton's second law in the form $g_n F = ma$.

Length units. The meter (m) is the SI base unit of length. The use of special names for decimal submultiples of the meter should be avoided, and units formed by attaching appropriate SI prefixes to the meter should be used instead. Thus the micron (μ), which was defined as 10^{-6} m, should be replaced by the micrometer (μm), which has the same value; and the millimicron (mμ), which was defined as 10^{-9} m, should be replaced by nanometer (nm). The fermi, which was defined as 10^{-15} m, and was used to measure nuclear distances, should be replaced by the femtometer (fm), which has the same value.

The angstrom (Å) is equal to 10^{-10} m. Although it has been accepted for temporary use with the SI, it is preferable to replace this unit with the nanometer, using the relation 1 Å = 0.1 nm.

The nautical mile (nmi), equal to 1852 m, has been accepted for temporary use with the SI in navigation. (This unit approximates the length of 1 minute of arc of a great circle on the surface of the Earth.) *See* NAVIGATION.

The foot (ft) is, as discussed above, the unit of length in the British systems of units, and it is also

TABLE 2. Gravitational systems of units

Quantity	Defining relation	Units		
		mks	cgs	British
Length (L)	Fundamental	m	cm	ft
Force (F)	Fundamental	kilogram force (kgf)	gram force (gf)	pound force (lbf)
Time (t)	Fundamental	s	s	s
Mass (m)	$m = F/a$	kgf · s^2/m	gf · s^2/cm	1 slug = 1 lbf · s^2/ft
Velocity (v)	$v = L/t$	m/s	cm/s	ft/s
Acceleration (a)	$a = v/t$	m/s^2	cm/s^2	ft/s^2
Work (W)	$W = FL$	m · kgf	cm · gf	foot-pound force (ft · lbf)
Power (P)	$P = W/t$	m · kgf/s	cm · gf/s	ft · lbf/s
Momentum (p)	$p = mv$	kgf · s	gf · s	slug · ft/s

in customary use in the United States. Since 1959 the foot has been defined as exactly 0.3048 m. The yard (yd) is defined as exactly 3 ft or 0.9144 m. *See* LENGTH; MEASURE.

Units used to measure x-ray wavelengths. Relative measurements of x-ray wavelengths can be made to a higher accuracy than absolute measurements. That is, the ratio of two x-ray wavelengths can be determined with a higher accuracy than the ratio of either of them to the meter. The same situation holds for dimensions of crystal lattices, which are derived from x-ray wavelengths by x-ray diffraction experiments. For this reason, x-ray wavelengths and dimensions of crystal lattices have been expressed in units that are defined in terms of a standard x-ray wavelength or crystal lattice dimension. *See* X-RAY CRYSTALLOGRAPHY; X-RAYS.

Before 1965, most x-ray wavelengths were expressed in terms of the X-unit, which is approximately 10^{-13} m. The grating constant of calcite was defined to be exactly 3029.04 X-units. Subsequent absolute measurements of x-ray wavelengths with ruled gratings indicated that the X-unit exceeds 10^{-13} m by about 2 parts per thousand. Furthermore, in practice, workers in the field began using definitions of the X-unit based on various x-ray wavelengths instead of the calcite grating definition, and subsequent precise measurements indicated that these wavelength standards differed from each other and from the calcite grating standard by as much as 20 parts per million.

The X-unit has been superseded by the A* unit, introduced by J. A. Bearden in 1965, which is based on the tungsten $K\alpha_1$ line as a standard. The peak of this line is defined as exactly 0.2090100 A*. X-ray wavelength tables have been published in terms of this unit. At the time the A* unit was defined, it was thought to equal 10^{-10} m (the angstrom unit, Å) to within 5 parts per million, but the A* unit is now believed to be 20 ± 5 parts per million larger than 10^{-10} m.

Units used in astronomy. Special units whose values are obtained experimentally are used in astronomy. For their definitions *see* ASTRONOMICAL UNIT; LIGHT-YEAR; PARSEC.

The astronomical unit and parsec are accepted for use with the SI. The parsec rather than the light-year is used in technical literature.

Area units. The square meter (m^2), the SI unit of area, is the area of a square with sides of length 1 m. Other area units are defined by forming squares of various length units in the same manner. The hectare (ha) is equal to 1 square hectometer (1 hm^2) or equivalently to 10^4 m^2. Its use with the SI is permitted for expressing land or water areas. *See* AREA.

Cross sections, which measure the probability of interaction between an atomic nucleus, atom, or molecule and an incident particle, have the dimensions of area, and the appropriate SI unit for expressing them is therefore the square meter. The barn (b), a unit of cross section equal to 10^{-28} m^2, has been accepted for temporary use with the SI. Typical nuclear reactions have cross sections ranging from millibarns to several thousand barns. A related quantity, which is connected with the probability that a reaction will emit radiations in a particular direction, is the differential cross section, which has the dimensions of barns per steradian. *See* NUCLEAR REACTION.

Units of volume. The cubic meter (m^3), the SI unit of volume, is the volume of a cube with sides of length 1 m. Other units of volume are defined by forming cubes of various length units in the same manner. The liter (symbol L in the United States) is equal to 1 cubic decimeter (1 dm^3), or equivalently to 10^{-3} m^3. It has been accepted for use with the SI for measuring volumes of liquids and gases.

Time units. The second (s) is the SI base unit of time. However, other units of time in customary use, such as the minute (1 min = 60 s), hour (1 h = 60 min), and day (1 d = 24 h), are acceptable for use with the SI. *See* TIME.

Frequency units. The hertz (Hz), the SI unit of frequency, is equal to 1 cycle per second. A periodic oscillation has a frequency of n hertz if it goes through n cycles in 1 s. Other units of frequency are defined by forming reciprocals of time units in the same manner. *See* FREQUENCY (WAVE MOTION).

Speed and velocity units. The meter per second (m/s), the SI unit of speed or velocity, is the magnitude of the constant velocity at which a body traverses 1 m in 1 s. Other speed and velocity units are defined by dividing a unit of length by a unit of time in the same manner. *See* SPEED; VELOCITY.

The knot (kn) is equal to 1 nautical mile per hour (1 nmi/h); it has been accepted for temporary use with the SI.

Acceleration units. The meter per second squared (m/s^2), the SI unit of acceleration, is the acceleration of a body whose velocity changes by 1 m/s in 1 s. Other units of acceleration are defined by dividing a unit of velocity by a unit of time in the same manner. *See* ACCELERATION.

The gal or galileo (symbol Gal) is equal to 1 cm/s^2, or equivalently to 10^{-2} m/s^2. This unit and its decimal submultiple the milligal (1 mGal = 10^{-3} Gal = 10^{-5} m/s^2) are employed in geodesy and geophysics to express the acceleration of gravity, and have been accepted for temporary use with the SI. *See* EARTH, GRAVITY FIELD OF; GEODESY.

Mass units. The kilogram (kg), the SI base unit of mass, is the only SI unit whose name, for historical reasons, contains a prefix. Names of decimal multiples and submultiples of the kilogram are formed by attaching prefixes to the word gram (g). Since 1 kg = 10^3 g, 1 g = 10^{-3} kg. The metric ton (t), which is equal to 10^3 kg or 1 megagram (Mg), is permitted in commercial usage of the SI.

The pound (lb), is, as discussed above, the unit of mass in the British absolute system, and is also in customary use in the United States. In 1959 the pound was defined to be exactly 0.45359237 kg.

The slug is, as discussed above, the unit of mass in the British gravitational system. By definition 1 pound force (lbf) acting on a body of mass 1 slug produces an acceleration of 1 foot per second

squared (1 ft/s^2). The slug is equal to approximately 32.174 lb or 14.594 kg. *See* MASS.

Force units. The newton (N), the SI unit of force, is the force which imparts an acceleration of 1 meter per second squared (1 m/s^2) to a body having a mass of 1 kg.

The dyne, the cgs absolute unit of force, is the force which imparts an acceleration of 1 centimeter per second squared (1 cm/s^2) to a body having a mass of 1 g. Since 1 cm/s^2 = 10^{-2} m/s^2, and 1 g = 10^{-3} kg, it follows that 1 dyne = 10^{-5} N.

The unit of force in the British absolute system is, as discussed above, the poundal (pdl), the force which imparts an acceleration of 1 foot per second squared (1 ft/s^2) when applied to a body of mass 1 lb. One poundal is approximately 0.13825 N.

As discussed above, the units of force in the mks gravitational, cgs gravitational, and British gravitational systems are the forces which impart an acceleration equal to the standard acceleration of gravity, g_n = 9.80665 m/s^2 ≅ 32.174 ft/s^2, when applied to bodies having masses of 1 kg, 1 g, and 1 lb, respectively. These units are named the kilogram force (kgf), gram force (gf), and pound force (lbf), respectively. Unfortunately, these units have also been called simply the kilogram, gram, and pound, giving rise to confusion with the mass units of the same name; in proper usage the suffix "force" should always be employed. The pound has also been called the pound mass (lbm) to further distinguish mass and force units. One pound force is approximately 4.4482 N. *See* FORCE.

Pressure and stress units. The pascal (Pa), the SI unit of pressure and stress, is the pressure or stress of 1 newton per square meter (N/m^2). This is a rather small unit for most practical purposes; for example, atmospheric pressure is approximately 10^5 Pa. Thus, most pressures are most readily expressed in decimal multiples of the pascal formed by attaching the appropriate SI prefix.

The dyne per square centimeter (dyne/cm^2), the cgs absolute unit of pressure, has sometimes been called the barye, but this name is uncommon. Other units of pressure can also be formed by dividing various units of force by various units of area, such as the pound force per square inch (lbf/in.^{2}, frequently abbreviated psi).

Pressure has been frequently expressed in terms of the bar and its decimal submultiples, where 1 bar = 10^6 dynes/cm^2 = 10^5 Pa. A reference level of 1 microbar (1 μbar = 10^{-6} bar = 1 dyne/cm^2 = 0.1 Pa) is commonly used in the calibration of microphones, hydrophones, and loudspeakers. The millibar (1 mbar = 10^{-3} bar = 10^2 dynes/cm^2 = 10^2 Pa = 0.1 kPa) is commonly used in meteorology. The temporary use of the millibar with the SI has been allowed in order to permit meteorologists to communicate easily within their profession, but the kilopascal should be used in presenting meteorological data to the public. *See* SOUND PRESSURE.

Pressures are also frequently expressed in terms of the height of a column of either mercury or water which the pressure will support. This practice is very convenient in conjunction with the use of barometers or other instruments in which pressure is determined from such column heights. However, the pressure which supports a liquid column of given height depends on the acceleration due to gravity and on the density of the liquid (and, in turn, on its temperature), and these must therefore be specified for accurate work.

Two other units which have been frequently used for measuring pressure are the standard atmosphere and the torr. The standard atmosphere (atm) is exactly 101,325 Pa, which is approximately the average value of atmospheric pressure at sea level. The torr is exactly 1/760 atmosphere, or approximately 133.322 Pa. To within 1 part per million, it is equal to the pressure of a column of mercury of height 1 millimeter (1 mmHg) at a temperature of 0°C when the acceleration due to gravity has the standard value g_n = 9.80665 m/s^2. *See* ATMOSPHERE; PRESSURE; PRESSURE MEASUREMENT.

Energy and work units. The joule (J), the SI unit of energy or work, is the work done by a force of magnitude 1 newton when the point at which the force is applied is displaced 1 m in the direction of the force. Thus, joule is a short name for newton-meter (N · m) of energy or work. *See* ENERGY; WORK.

Units of energy or work in other systems are defined by forming the product of a unit of force and a unit of length in precisely the same manner as in the definition of the joule. Thus, the erg, the cgs absolute unit of energy or work, is the product of 1 dyne and 1 cm. Erg is a short name for dyne-centimeter of energy or work. Since 1 dyne = 10^{-5} N and 1 cm = 10^{-2} m, it follows that 1 erg = 10^{-7} J.

The foot-poundal (ft · pdl), the British absolute unit of energy or work, is the product of 1 poundal and 1 foot. The foot-pound, or, more properly, the foot-pound force (ft · lbf), the British gravitational unit of energy or work, is the product of 1 lbf and 1 ft. Foot-poundal and foot-pound are also names of units of torque, discussed below.

Power-time products. Sometimes energy is measured in units which are products of a unit of power and a unit of time. Since 1 watt (W) of power equals 1 joule per second (1 J/s), as discussed below, the joule is equivalent to 1 watt-second (1 W · s). In electrical power applicatons, energy is frequently measured in kilowatt-hours (kWh), where 1 kWh = (10^3 W) · (3600 s) = 3.6 × 10^6 J. *See* ELECTRICAL ENERGY MEASUREMENT.

Calorie and British thermal unit. The calorie and the British thermal unit were originally defined as the quantities of heat required to raise the temperature of a specified mass of water by a specified temperature. Usually these units have been used in connection with energy as heat, but they can also be used when referring to energy as work or energy in any other form. *See* CALORIMETRY; HEAT.

The calorie was originally defined as the quantity of heat required to raise the temperature of 1 g of air-free water 1°C under a constant pressure of 1 atm. However, the magnitude of the calorie, so defined, depends on the place on the Celsius temperature

scale at which the measurement is made. The 15°C calorie is based on the temperature interval from 14.5 to 15.5°C, but other temperature intervals have been used. At the Fifth International Conference on the Properties of Steam in London in 1956, the International (Steam) Table calorie was defined as exactly 4.1868 J; this is the type of calorie most frequently used in mechanical engineering. The thermochemical calorie, which has been used in thermochemistry in preference to the other types of calorie, is exactly 4.184 J.

Since the calorie is a relatively small unit, the kilocalorie (kcal), also called the large calorie and the kilogram-calorie, is used to designate 10^3 calories. The energy intake of the human body has been commonly expressed as the number of large calories which the eaten food will liberate as it passes through the body. Unfortunately, confusion arises because it is usually not indicated that kilocalories rather than calories are being used.

The British thermal unit (Btu) was originally defined as the quantity of heat required to raise the temperature of 1 lb of air-free water 1°F under a constant pressure of 1 atm. Again, the temperature interval must be specified; the 60° Btu, based on the interval from 59.5 to 60.5°F, has been frequently used.

The International Table Btu and the thermochemical Btu are both defined to be of such magnitude that the values of specific heat capacity of any substance are equal in size whether expressed in Btu per pound per degree Fahrenheit [Btu/(lb · °F)] or in calories per gram per degree Celsius [cal/(g · °C)], when the corresponding type of calorie is used. From this it follows that each type of Btu is equal to approximately 251.996 times the corresponding type of calorie. Then the International Table Btu is approximately 1055.056 J, and the thermochemical Btu is approximately 1054.350 J.

Electronvolt. The electronvolt (eV), whose value is experimentally determined, is frequently used to express the energies of atomic systems. It has been accepted for use with the SI. *See* ELECTRONVOLT.

Power units. The watt (W), the SI unit of power, is the power which gives rise to the production of energy at the rate of 1 joule per second (1 J/s). Other units of power can be defined by forming the ratio of a unit of energy to a unit of time in the same manner. *See* POWER.

The horsepower (hp) is equal to exactly 550 ft · lb/s, or approximately 745.700 W. It frequently has been employed to express the power generated by engines and machinery.

Torque units. The newton-meter (N · m), the SI unit is the magnitude of the torque produced by a force of 1 newton acting at a perpendicular distance of 1 m from a specified axis of rotation. The joule should never be used as a synonym for this unit: although the two units are both products of 1 newton and 1 m, the orientation of force and length is quite different in the two cases.

Units of torque in other systems are defined by forming the product of a unit of force and a unit of length in precisely the same manner as in the definition of the newton-meter. Thus the dyne-centimeter (dyne · cm), the cgs absolute unit of torque, is the product of 1 dyne and 1 cm; the erg should never be used as a synonym for this unit.

The foot-poundal (ft · pdl), the British absolute unit of torque, is the product of 1 poundal and 1 foot. The foot-pound (ft · lbf), the British gravitational unit of torque, is the product of 1 lbf and 1 ft. These units are sometimes called the poundal-foot (pdl · ft) and pound-foot (lbf · ft) to distinguish them from the units of energy or work.

The SI unit of torque does work of 1 N · m on a body which rotates through 1 rad in the direction of the torque. Some standards of metric practice therefore suggest that the SI unit of torque can be designated newton-meter per radian (N · m/rad) in order to further distinguish it from the unit of energy or work. The names of other torque units should then be similarly modified. *See* ROTATIONAL MOTION; TORQUE.

Viscosity units. For units of dynamic viscosity and kinematic viscosity *see* VISCOSITY.

Permeability units. The darcy is a commonly used unit of permeability to fluid flow. The permeability of a rock, brick, or other porous substance is 1 darcy if 1 cm^3 of a fluid of 1-centipoise viscosity will flow through a section 1 cm thick and of $1-cm^2$ cross section in 1 s at a pressure difference of 1 atm. In core analysis and other measurements of petroleum-bearing rock, the unit used is the millidarcy, which is 10^{-3} darcy. *See* PETROLEUM RESERVOIR ENGINEERING; SOIL MECHANICS.

The SI unit of permeability is the permeability of a porous substance such that 1 m^3 of a fluid of viscosity 1 Pa · s will flow through a section 1 m thick and of 1-m^2 cross section in 1 s at a pressure difference of 1 Pa. This unit has no special name, but has the dimensions of 1 m^2. The SI unit of permeability is exactly 1.01325×10^{12} darcy, and 1 darcy $\equiv 0.987 \times 10^{-12}$ m^2.

Electrical Units

For a general discussion of electrical units, including the SI or mks system, three cgs systems [electrostatic system of units (esu), electromagnetic system of units (emu), and gaussian system], and definitions of the SI units ampere (A), volt (V), ohm (Ω), coulomb (C), farad (F), henry (H), and weber (Wb) *see* ELECTRICAL UNITS AND STANDARDS

This section discusses some additional SI units and some units in the cgs electromagnetic system which are frequently encountered in scientific literature in spite of the fact that their use has been discouraged.

Siemens. The siemens (S), the SI unit of electrical conductance, is the electrical conductance of a conductor in which a current of 1 ampere is produced by an electric potential difference of 1 volt. The conductance G is defined by the equation $I = GV$, where I is the current in amperes, V is the potential difference in volts, and G the conductance in siemens. The conductance of an electrical conductor in siemens is the reciprocal of its resistance in ohms. *See* CONDUCTANCE; ELECTRICAL RESISTANCE.

The siemens was formerly called the mho (℧) to illustrate the fact that the unit is the reciprocal of the ohm.

Abampere. The abampere (abA), the cgs electromagnetic unit of current, is that current which, if maintained in two straight, parallel conductors of infinite length, of negligible circular cross section, and placed 1 cm apart in vacuum, would produce between these conductors a force equal to 2 dynes per centimeter of length. *See* ELECTRIC CURRENT.

The abampere is equal to exactly 10 A. That a current of 10 A satisfies the above definition can be seen from the following argument: If two straight, parallel conductors of infinite length, separated by distance r, carry currents I_1 and I_2, then the force F on segment of one of the conductors of length l is given by Eq. (5). The constant μ_0 is the permeability of vac-

$$F = \frac{\mu_0 I_1 I_2 l}{2\pi r} \tag{5}$$

uum, $4\pi \times 10^{-7}$ newton per ampere squared. Thus, when the current in each wire is 1 A and the distance between the conductors is $r = 1$ m, the force between them is 2×10^{-7} newton per meter of length, as in the definition of the ampere. When $I_1 = I_2 = 10$ A, and $l = r = 1$ cm $= 10^{-2}$ m, the force between the conductors, given by Eq. (6), is the force speci-

$$F = \frac{(4\pi \times 10^{-7})\,(10)\,(10)\,(10^{-2})}{2\pi(10^{-2})} \text{ newton}$$

$$= 2 \times 10^{-5} \text{newton} = 2 \text{ dynes} \tag{6}$$

fied in the definition of the abampere. *See* AMPÈRE'S LAW; BIOT-SAVART LAW.

Abvolt. The abvolt (abV), the cgs electromagnetic unit of electrical potential difference and electromotive force, is the difference of electrical potential between two points of a conductor carrying a constant current of 1 abA, when the power dissipated between these points is equal to 1 erg per second. Then 1 abV $= 10^{-8}$ V, as can be seen from Eq. (7).

$$1 \text{ abV} = \frac{1 \text{ erg/s}}{1 \text{ abA}} = \frac{10^{-7} \text{ J/s}}{10 \text{ A}}$$

$$= 10^{-8} \frac{\text{W}}{\text{A}} = 10^{-8} \text{ V} \tag{7}$$

The abampere and abvolt are not frequently used, but they figure in the definitions of the maxwell, gauss, oersted, and gilbert, given below.

Maxwell. The maxwell (Mx), the cgs electromagnetic unit of magnetic flux, is the magnetic flux which, linking a circuit of one turn, produces in it an electromotive force of 1 abV as it is reduced to zero in 1 s. Then 1 maxwell $= 10^{-8}$ weber, as can be seen from Eq. (8).

$$1 \text{ Mx} = 1 \text{ abV} \cdot \text{s} = 10^{-8} \text{ V} \cdot \text{s} = 10^{-8} \text{ Wb} \tag{8}$$

Units of magnetic flux density. The tesla (T), the SI unit of magnetic flux density (also called magnetic induction), is a magnetic flux density of 1 weber per square meter (1 Wb/m^2). The gauss (Gs), the cgs electromagnetic unit of magnetic flux density, is a magnetic flux density of 1 maxwell per square centimeter (1 Mx/cm^2). Then 1 gauss $= 10^{-4}$ tesla, as can be seen from Eq. (9).

$$1\text{Gs} = \frac{1 \text{ Mx}}{(1 \text{ cm})^2} = \frac{10^{-8} \text{ Wb}}{(10^{-2} \text{ m})^2}$$

$$= 10^{-4} \frac{\text{Wb}}{\text{m}^2} = 10^{-4} \text{ T} \tag{9}$$

Alternative equivalent definitions for the tesla and gauss follow from Eq. (10) for the magnetic induc-

$$B = \frac{F}{Il \sin\theta} \tag{10}$$

tion B, where F is the force exerted on a current element of length l and current I making angle θ with the magnetic induction vector. This equation is valid whether quantities are expressed in SI or in cgs electromagnetic units. Thus, an alternative definition for the tesla is the constant magnetic induction which exerts a force of 1 newton on a straight wire of length 1 m perpendicular to the magnetic induction vector and carrying a current of 1 ampere; that is, 1 T = 1 N/(A · m). The gauss can be defined, in precisely the same manner, as 1 dyne per abampere-centimeter [1 Gs = 1 dyne/(abA · cm)].

Units of magnetic field strength. The SI unit of magnetic field strength is 1 ampere per meter (1 A/m), which is the magnetic field strength at a distance of 1 m from a straight conductor of infinite length and negligible circular cross section which carries a current of 2π A. (Other geometric configurations can be used to define this unit; the choice is arbitrary.) This definition is based on the definition, in the SI, of the magnetic field strength H_{SI}. At a distance r from a long straight conductor carrying current I, H_{SI} is given by Eq. (11). The left-hand side of this equation

$$2\pi r H_{\text{SI}} = I \tag{11}$$

is the line integral of H_{SI} around a circular path, all of whose points are at distance r from the conductor. Substituting $r = 1$ m and $I = 2\pi$ A gives $H_{\text{SI}} = 1$ A/m.

The SI unit has sometimes been called the ampere-turn per meter because Eq. (11) can be modified to Eq. (12), where i is the current in each turn of a

$$2\pi r H_{\text{SI}} = ni \tag{12}$$

coil, part of which forms the long straight conductor in question, and n is the number of turns in the coil. Substituting $r = 1$ m, $i = 2\pi$ A, and $n = 1$ turn in Eq. (12) gives $H_{\text{SI}} = 1$ A · turn/m. However, the name ampere per meter has been adopted by the CGPM.

The oersted (Oe), the cgs electromagnetic unit of magnetic field strength, is the magnetic field strength at a distance of 1 cm from a straight conductor of infinite length and negligible circular cross section which carries a current of 0.5 abA. Again, this definition is based on the definition of magnetic field strength. In the cgs electromagnetic system the magnetic field strength H_{cmu} at a distance r from a long straight conductor carrying current I is given by

Eq. (13), which differs from Eq. (11) by the factor

$$2\pi r H_{\text{emu}} = 4\pi I \tag{13}$$

4π on the right-hand side. Substituting $r = 1$ cm and $I = 0.5$ abA in this equation gives $H_{\text{emu}} = 1$ abA/cm. Thus, 1 Oe = 1 abA/cm in the cgs electromagnetic system, but the method by which the quotient is formed is different from that in the SI unit, being based on Eq. (13) rather than Eq. (11).

When the magnetic field strength in cgs electromagnetic units is 1 Oe, the magnetic field strength in SI units is found by substituting in Eq. (11) $I = 0.5$ abA and $r = 1$ cm, which gives Eq. (14). Thus

$$H_{\text{SI}} = \frac{0.5 \text{ abA}}{2\pi(1 \text{ cm})} = \frac{0.5(10 \text{ A})}{2\pi(10^{-2} \text{ m})} = \frac{10^3}{4\pi} \frac{\text{A}}{\text{m}} \tag{14}$$

1 Oe corresponds to $(10^3/4\pi)$ A/m $\cong$ 79.577 A/m.

Units of magnetic potential and mmf. The ampere serves as the SI unit of magnetic potential difference and magnetomotive force (mmf), as well as the unit of current. In the SI system, the magnetomotive force around a closed path equals the current passing through a surface enclosed by the path. Thus, 1 A is the magnetomotive force around a closed path when a current of 1 A passes through an enclosed surface. The SI unit has also been called the ampere-turn, because the magnetomotive force around a path that loops around a current coil equals the product of the number of turns in the coil and the current in each turn; however, the name ampere has been adopted by the CGPM.

The gilbert (Gb), the cgs electromagnetic unit of magnetic potential difference and magnetomotive force, is the magnetomotive force around a closed path enclosing a surface through which flows a current of $[1/(4\pi)]$ abA. This definition is based on the fact that in the cgs electromagnetic system the magnetomotive force around a closed path is 4π times the current passing through an enclosed surface. When this current is $[1/(4\pi)]$ abA, the magnetomotive force is therefore 1 abA in cgs electromagnetic units, but the unit is given the special name gilbert. When the magnetomotive force in cgs electromagnetic units is 1 gilbert, the magnetomotive force in SI units is $[1/(4\pi)]$ abA = $[10/(4\pi)]$ A. Thus, 1 Gb corresponds to $[10/(4\pi)]$ A $\cong$ 0.79577 A. *See* MAGNETOMOTIVE FORCE.

Photometric Units

Photometric units involve a new base quantity, luminous intensity. For the definition of the candela (cd), the SI unit of luminous intensity, *see* PHOTOMETRY; PHYSICAL MEASUREMENT.

For a general discussion of photometric units, including units of illuminance (illumination) and luminance, and in particular the SI units lux (lx) and candela per square meter (cd/m^2), *see* ILLUMINATION; LUMINANCE.

This section gives an explicit definition of the lumen and discusses units of luminous energy.

Lumen. The lumen (lm), the SI unit of luminous flux, is the luminous flux emitted within a unit solid angle (1 steradian) by a point source having a uniform intensity of 1 candela. It follows, therefore, that a light source having an intensity of 1 candela in every direction will be emitting a total luminous flux of 4π lumens.

The lumen is also equal to the luminous flux received on a unit surface, all points of which are at a unit distance from a point source having a uniform intensity of 1 candela.

The output of light sources is given in lumens. *See* LUMINOUS FLUX.

Luminous energy units. The lumen-second (lm · s), the SI unit of luminous energy (also called quantity of light), is the luminous energy radiated or received over a period of 1 s by a luminous flux of 1 lumen. This unit is also called the talbot. Other units of luminous energy include the lumen-hour (1 lm · h = 3600 lm · s) and the million-lumen-hour. *See* LUMINOUS ENERGY.

Radiation Units

Certain quantities and units are used particularly in the area of ionizing radiation. The special units curie, roentgen, rad, and rem, which were previously adopted for use in this area, are not coherent with the SI, but their temporary use with the SI has been approved while the transition to SI units takes place. The CGPM, acting on proposals of the International Commission on Radiation Units and Measurements (ICRU) and the International Commission on Radiological Protection (ICRP), has adopted the special names becquerel, gray, and sievert for the SI derived units of activity, absorbed dose, and dose equivalent.

Activity units. The becquerel (Bq), the SI unit of activity (radioactive disintegration rate), is the activity of a radionuclide decaying at the rate of one spontaneous nuclear transition per second. Thus 1 Bq = 1 s^{-1}.

The curie (Ci), the special unit of activity, is equal to 3.7×10^{10} Bq. (This unit was originally chosen to approximate the activity of 1 g of radium-226.) *See* RADIOACTIVITY.

Exposure units. The SI unit of exposure to ionizing radiation, 1 coulomb per kilogram (1 C/kg), is the amount of electromagnetic radiation (x-radiation or gamma radiation) which in 1 kg of pure dry air produces ion pairs carrying 1 coulomb of charge of either sign. (The ionization arising from the absorption of bremsstrahlung emitted by electrons is not to be included in measuring the charge). *See* BREMSSTRAHLUNG.

The roentgen (R), the special unit of exposure, is equal to 2.58×10^{-4} C/kg. (This is equivalent to 1 cgs electrostatic unit of charge in 1.293 mg of air, which is the mass of 1 cm^3 of air at temperature 0°C and pressure 1 atm.)

Absorbed dose and kerma units. The gray (Gy), the SI unit of absorbed dose, is the absorbed dose when the energy per unit mass imparted to matter by ionizing radiation is 1 joule per kilogram (1 J/kg).

The rad (rd), the special unit of absorbed dose, is equal to 10^{-2} Gy. In air, the number of rad per

roentgen is 0.877, and this is also approximately true in soft tissue.

Kerma (an acronym for kinetic energy released in matter) is related to absorbed dose and is measured in the same units. The kerma is the sum of the initial kinetic energies of all the charged ionizing particles liberated by uncharged ionizing particles in an element of matter divided by the mass of that element. Except for a correction due to energy lost to bremsstrahlung, the exposure is the ionization equivalent of the air kerma.

Dose equivalent units. Different types of radiation cause slightly different effects in biological tissue. For this reason, a weighted absorbed dose called the dose equivalent is used in comparing the effects of radiation on living systems. The dose equivalent is the product of the absorbed dose and various dimensionless modifying factors stipulated by the ICRP. The chief such factor, the quality factor, depends on the linear energy transfer, and this, in turn, depends on the kind of incident radiation and its energy. For electromagnetic radiation and electrons the quality factor equals 1; for heavier charged particles it is greater than 1. *See* LINEAR ENERGY TRANSFER (BIOLOGY).

The sievert (Sv), the SI unit of dose equivalent, is the dose equivalent when the absorbed dose of ionizing radiation multiplied by the stipulated dimensionless factors is 1 joule per kilogram (1 J/kg).

The rem, the special unit of dose equivalent, is equal to 10^{-2} Sv. *See* ENVIRONMENTAL RADIOACTIVITY; RADIATION BIOLOGY; RADIATION INJURY (BIOLOGY).

Unit of Catalytic Activity

In 1999, the CGPM adopted the katal (kat) as the SI unit of catalytic activity, equal to 1 mole per second (1 mol/s). In its resolution adopting the unit, the CGPM cited the importance for human health and safety of facilitating the use of SI units in the fields of medicine and biochemistry, and noted that a non-SI unit of catalytic activity called "unit," symbol U, equal to 1 micromole per minute (1μmol/min), which is not coherent with SI, has been in widespread use in these fields.

Other Units

Logarithmic measures may be used with the SI. *See* BEL; DECIBEL; NEPER; PH; VOLUME UNIT (VU). For units of loudness (sone, phon) *see* LOUDNESS. For units of sound absorption by surfaces *see* ARCHITECTURAL ACOUSTICS. For units of acoustic frequency and pitch *see* MUSICAL ACOUSTICS; PITCH. For quantities and units pertaining to sound transmission *see* SOUND. For temperature scales *see* TEMPERATURE. For units measuring quantities in chemistry *see* ATOMIC MASS UNIT; ATOMIC WEIGHT; CONCENTRATION SCALES; ELECTROCHEMICAL EQUIVALENT; GRAM-MOLECULAR WEIGHT; MOLECULAR WEIGHT; RELATIVE ATOMIC MASS; RELATIVE MOLECULAR MASS. For quantities and units pertaining to nuclear reactors *see* REACTOR PHYSICS. For units of information content *see* BIT; INFORMATION THEORY.

Jonathan F. Weil

Bibliography. Institute of Electrical and Electronics Engineers, *Standard for Use of the International System of Units (SI): The Modern Metric System*, IEEE/ASTM Std. SI 10-2002, 2002; International Commission on Radiation Units and Measurements, *Fundamental Quantities and Units for Ionizing Radiation*, ICRU Rep. 60, 1998; International Commission on Radiation Units and Measurements, *Quantities and Units in Radiation Protection Dosimetry*, ICRU Rep. 51, 1993; International Organization for Standardization, *Quantities and Units*, 1993; B. N. Taylor, *Guide for the Use of the International System of Units (SI)*, NIST Spec. Pub. 811, 1995; B. N. Taylor (ed.), *The International System of Units (SI)*, NIST Spec. Pub. 330, revised 2001; U. S. Metric Association, *Guide to the Use of the Metric System*, 15th ed., 2002.

Universal joint

A linkage that transmits rotation between two shafts whose axes are coplanar but not coinciding. The universal joint is used in almost every class of machinery: machine tools, instruments, control devices, and, most familiarly, automobiles.

Hooke's joint. A simple universal joint, known in English-speaking countries as Hooke's joint and in continental Europe as a Cardan joint, is shown in **Fig. 1***a*. It consists of two yokes *A* and *B* (Fig. 1*b*) attached to their respective shafts and connected by means of spider *S*. Angle ϕ between the shafts may have any value up to approximately 35°, if angular velocity is moderate when angle ϕ is large. Although shaft *B* must make one revolution for each revolution of shaft *A*, the instantaneous angular displacement of shaft *B* is the same as that of shaft *A* only at the end of each 90° of shaft rotation. Thus, only at four positions during each revolution is angular velocity ω_B of shaft *B* the same as angular velocity ω_A of shaft *A*. Three curves for designated values of ϕ are plotted in **Fig. 2**. These curves show deviation of ω_B from a constant ω_A as shaft *A* is turned through 180°. *See* FOUR-BAR LINKAGE.

Double Hooke's joint. The variation in angular displacement and angular velocity between driving and driven shafts, which is objectionable in many mechanisms, can be eliminated by using two Hooke's joints, with an intermediate shaft (Fig. 1*c*). This arrangement is conventional for an automobile drive shaft. The axes of the driving and driven shafts need not intersect; however, it is necessary that the axes y and y' of the two yokes attached to intermediate shaft *b* lie respectively in planes containing the axes of adjoining shafts (*a*, y, and *b* in same plane; *b*, y', and *c* in same plane), and that angle β between the driving and intermediate shafts equal angle β' between the intermediate and driven shafts.

Bendix-Weiss joint. Two intersecting, thin, bent shafts (**Fig. 3***a*) with plane *Z* through the point of contact bisecting the angle between the shafts and perpendicular to the plane containing the axes of the shafts will maintain the constant angular velocity.

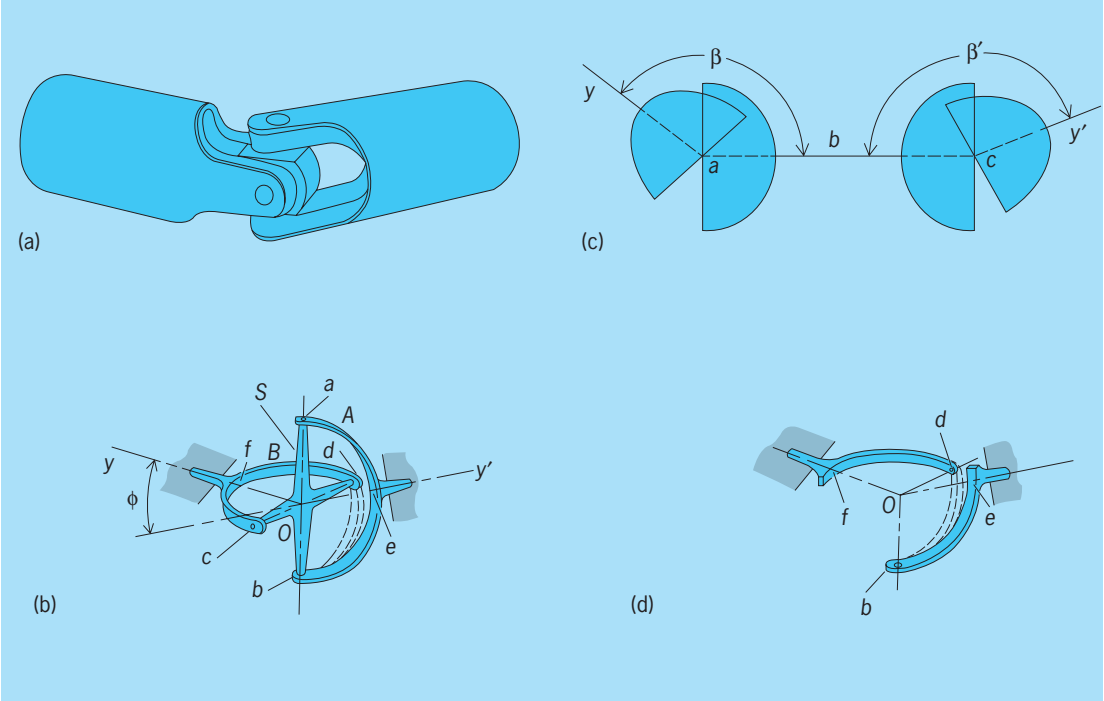


Fig. 1. Universal joints. (*a*) Simple. (*b*) Yoke and spider. (*c*) Double. (*d*) Four-bar conic linkage equivalent of yoke and spider. (*After C. W. Ham, E. J. Crane, and W. L. Rogers, Mechanics of Machinery, McGraw-Hill, 1958*)

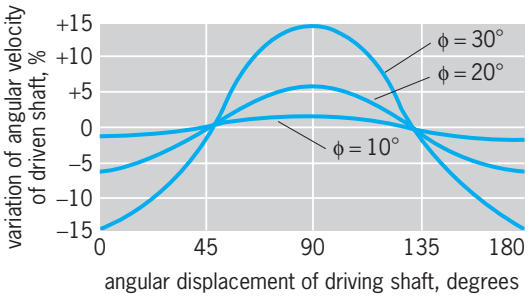


Fig. 2. Variation in the angular velocity of the driven shaft of a simple universal joint for the constant angular velocity of the driving shaft.

One practical application of this principle, in which constant angular velocity is transmitted through a single universal joint, is the joint developed by Carl Weiss and illustrated in Fig. 3*b*. Four large balls are transmitting elements, while a center ball acts as a spacer. The transmitting balls must lie in plane *Z*, as explained above. By means of milled grooves in the yoke attached to each shaft, the balls are maintained at all times in plane *Z*. This joint (**Fig. 4**) is used for a front-wheel automotive drive because it transmits unvarying angular displacement even when steering the vehicle requires a varying angle between driving and driven shafts.

Spherical four-bar linkage. The universal joint is a spatial, spherical four-bar linkage. If only half of each fork is considered as *eb* of *A* and *fd* of *B* (Fig. 1*b*) and these are assumed to be connected by spherical link *db* equal to the fixed distance between the two adjacent points of spider *S*, a four-bar conic linkage is produced (Fig. 1*d*) in which the axes of all the turning pairs intersect in *O*. With this arrangement the fork could be omitted, and there would be a kinematic equivalent of the original mechanism.

Driver and follower make complete revolutions in the same time, but the velocity ratio is not constant

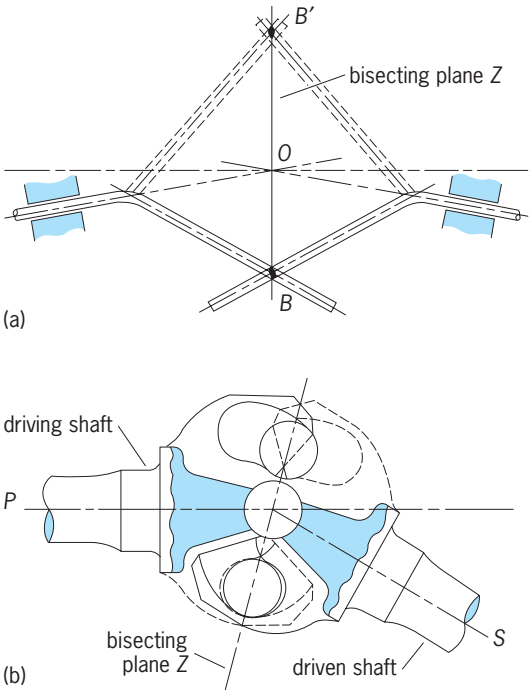


Fig. 3. Interacting shafts in sliding contact transmit constant angular velocity. (*a*) Basic configuration (*after R. T. Hinkle, Kinematics of Machines, Prentice-Hall, 1952*). (*b*) Cross section of ball-bearing adaptation.

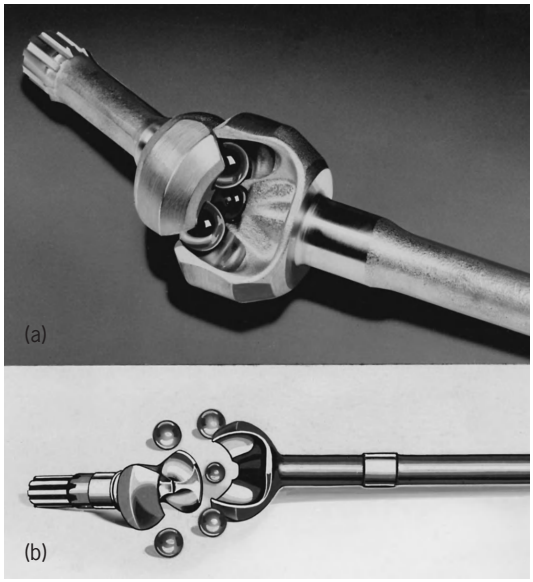


Fig. 4. Constant-velocity universal joint sometimes called a Rzeppa joint. (*a*) Partially separated cutaway. (*b*) Disassembled joint showing the arrangement; the right-hand member has been cut away. (*Bendix Aviation Corp.*)

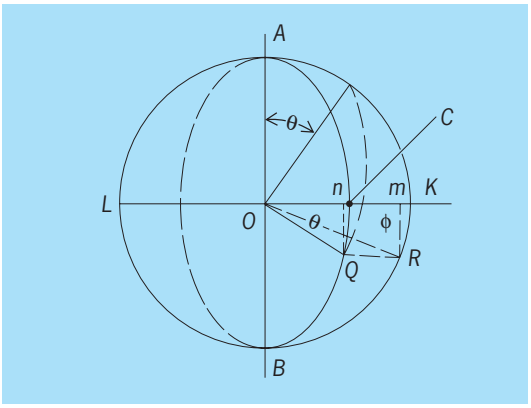


Fig. 5. Geometry of motion for spherical joint of Fig. 1*d*.

throughout the revolution as can be shown by analysis. In **Fig. 5**, COQ is the projection of the real angle described by the follower. The real component of the motion Qn of the follower is in a direction parallel to AB, and line AB is the intersection of the planes of the driver's and the follower's path. The true angle ϕ described by the follower (while the driver describes the angle θ) can be found by revolving OQ about AB as an axis into the plane of circle $AKBL$. Then OR is the true length of OQ, and ROK, designated ϕ, equals the true angle that is projected as COQ, designated θ. With this notation $\tan \phi = Rm/Om$ and $\tan \phi = On/On$. But $Qn = Rm$; hence $\tan \theta/\tan \phi = Om/On = OK/OC = 1/\cos \beta$.

Douglas P. Adams

Bibliography. A. S. Hall, Jr., *Kinematics and Linkage Design*, 1986; H. H. Mabie and C. F. Reinholtz, *Mechanisms and Dynamics of Machinery*, 4th ed., 1987; G. H. Martin, *Kinematics and Dynamics of Machines*, 2d ed., 1982.

Universal motor

A series motor built to operate on either alternating current (ac) or direct current (dc). It is normally designed for capacities less than 1 hp (0.75 kW). It is usually operated at high speed, 3500 revolutions/min loaded and 8000 to 10,000 revolutions/min unloaded. For lower speeds, reduction gears are often employed, as in the case of electric hand drills or food mixers. As in all series motors, the rotor speed increases as the load decreases, and the no-load speed is limited only by friction and windage. To obtain more constant speed with variations in load, a centrifugal governor may be used to switch in or out a small resistor in series with the armature as in **Fig. 1**.

If an alternating current is applied to any dc series motor, the motor would still rotate. Since the current is reversed simultaneously in the armature and the field, the torque would pulsate but would not reverse direction. However, a universal motor designed to operate on ac should have certain modifications: laminated cores to avoid excessive eddy currents, fewer turns in the field coils than in a dc motor, and more poles and usually more commutator segments. *See* CORE LOSS; DIRECT-CURRENT MOTOR.

The series ac motor is an alternating-current commutator motor which has great flexibility of performance. It can be operated over a wide range of speeds and is readily controllable. The series ac commutator motor is in many respects similar to the dc series motor and the universal motor.

The ac series motor, like the dc series motor and the repulsion motor, consists fundamentally of these windings or their equivalent: (1) rotating armature winding, (2) stationary field winding, and (3) compensating winding (**Fig. 2**).

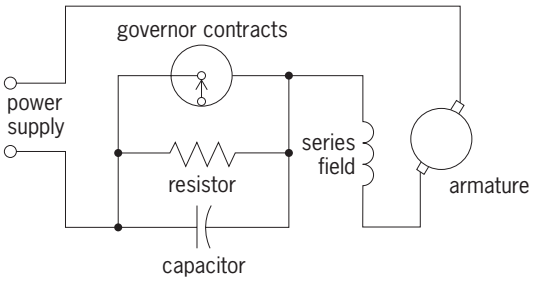


Fig. 1. Universal motor diagram.

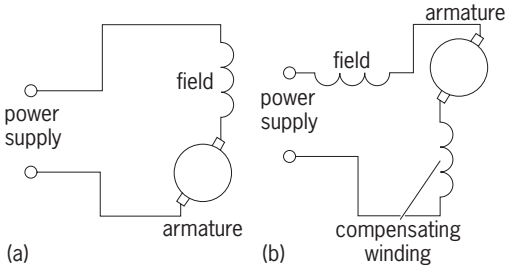


Fig. 2. Series motor diagram. (*a*) Without compensating winding. (*b*) With compensating winding.

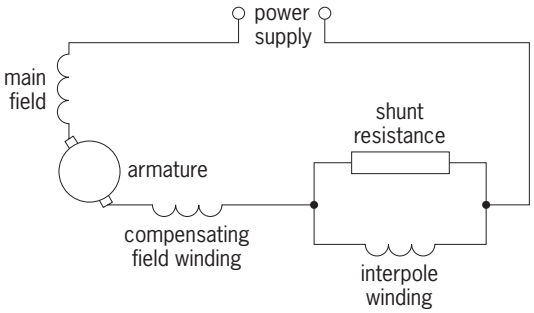


Fig. 3. Single-phase traction motor circuit diagram. (*General Electric Co.*)

A major problem in larger (up to 1000 hp or 750 kW) ac series motors is in commutation. Because of the transformer action between the field and armature coils, voltage is produced in the armature coils which are short-circuited by the brushes as the commutator bars pass under them. The coils which are short-circuited act like a short-circuited secondary of a static transformer. The resulting large currents are interrupted as the bars pass the brushes, causing bad sparking. In addition, these induced currents reduce the magnetic flux of the field and reduce the torque of the motor. Interpoles shunted with noninductive resistance are required on ac series motors as in **Fig. 3**. *See* COMMUTATION.

The single-phase commutator motor usually has a large number of turns in the armature winding, more commutator segments, and a small number of turns in the field winding, as compared with the dc motor, which is designed for relatively strong field and weaker armature. The ac motor usually has more poles and operates at a lower voltage than its dc counterpart.

Irving L. Kosow

Bibliography. D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 2000; A. E. Fitzgerald, C. Kingsley, and S. D. Umans, *Electric Machinery*, 5th ed., 1990; G. McPherson and R. D. Laramore, *An Introduction to Electrical Machines and Transformers*, 2d ed., 1990; S. A. Nasar, *Electric Machines and Power Systems*, 1995.

Universe

The universe comprises everything in existence, including all matter and energy, and the enormous volume which contains them. The observable universe currently spans about 2.6×10^{23} km (1.6×10^{23} mi), and contains approximately 2.4×10^{52} kg (5.2×10^{52} lb) of matter, yielding an average density equivalent to a few atoms per cubic meter. Most of the universe, then, is empty space; the matter is distributed thinly throughout, forming objects and structures at a variety of different sizes. The study of this matter and energy, and its distribution, composition, and origin, is what constitutes the sciences of astronomy and cosmology.

Constituents

This article will start the cosmic survey with the more familiar physical objects—atoms, stars, planets, and galaxies—following a sequence of increasing size, building up to the largest structures in the universe. A number of lesser-known and less tangible entities, such as energy, exotic particles, and dark matter, will complete the survey. With this synopsis of the constituents of the universe in hand, this article will address the questions of where they all came from and what their ultimate fate will be.

Ordinary units of measure, such as miles, kilometers, pounds, and kilograms, are usually not convenient for dealing with the enormous sizes and masses encountered in astronomy, so astronomers have devised other units. The standard length scale in astronomy is the parsec (pc), equal to 3.1×10^{13} km (1.9×10^{13} mi). For larger distances, kiloparsecs (kpc = 1000 pc) and megaparsecs (Mpc = 1,000,000 pc) are used. The light-year, about one-third of a parsec, often appears in popular astronomy texts, but is less common at the professional level. The mass of the Sun, 2.0×10^{30} kg (4.4×10^{30} lb), is used as a standard mass unit. Temperatures are measured using the Kelvin scale. A kelvin (K) is the same as a Celsius degree, but the Kelvin scale starts from absolute zero (-273°C) instead of the freezing point of water. To recast the figures quoted above, then, the observable universe has a diameter of roughly 8400 Mpc and contains 4.4×10^{22} solar masses of material.

Baryonic matter. Most of the matter encountered in everyday life is in the form of atoms. An atom consists of a positively charged nucleus of protons and neutrons, surrounded by clouds or shells of negatively charged electrons. The protons and neutrons are responsible for most of the mass of the atom. Since both protons and neutrons belong to a class of subatomic particles known as baryons, this ordinary form of atomic matter is called baryonic matter by astronomers. The number of protons in an atom's nucleus determines many of its properties, that is, which of the roughly 100 chemical elements that atom represents. *See* BARYON; ELEMENT (CHEMISTRY); ELEMENTARY PARTICLE; ELEMENTS, GEOCHEMICAL DISTRIBUTION OF; PERIODIC TABLE.

A large fraction of the visible matter elsewhere in the universe—planets, stars, nebulae, galaxies—is also baryonic in nature, but the relative proportions of the chemical elements are very different from here on Earth. Hydrogen is by far the most abundant element in the universe, representing 75% of the total baryonic mass. Helium is also plentiful, at about 23% of the total mass. All the heavier elements, collectively referred to as metals (regardless of their solidity or conductivity), make up the remaining 2%, with carbon, nitrogen, oxygen, neon, and iron as the primary constituents. The relative chemical proportions observed on the surface of the Sun (the solar abundances) are used as a reference point for most other abundance measurements (see **table**). The exact proportions present in any given

Solar abundances of the most common elements*

Element	Symbol	Atomic number	Mass fraction	Number fraction
Hydrogen	H	1	0.73	0.92
Helium	He	2	0.25	0.078
Oxygen	O	8	0.0077	0.00061
Carbon	C	6	0.0029	0.00030
Iron	Fe	26	0.0016	0.000037
Neon	Ne	10	0.0012	0.000077
Nitrogen	N	7	0.00095	0.000084
Silicon	Si	14	0.00069	0.000030
Magnesium	Mg	12	0.00047	0.000024
Sulfur	S	16	0.00038	0.000015
Argon	Ar	18	0.00018	0.0000058
Nickel	Ni	28	0.000086	0.0000018
Calcium	Ca	20	0.000058	0.0000018
Aluminum	Al	13	0.000044	0.0000023
Sodium	Na	11	0.000030	0.0000016
Chromium	Cr	24	0.000027	0.00000065
Chlorine	Cl	17	0.000012	0.00000037
Manganese	Mn	25	0.000010	0.00000023
Phosphorus	P	15	0.0000075	0.00000030
Cobalt	Co	27	0.0000058	0.00000012

*The proportion of each element is listed both by mass (fraction of the total mass) and by number (fraction of the total number of atoms).
SOURCE: Data after C. W. Allen, *Astrophysical Quantities*, 3d ed., Althone Press, London, 1991.

astronomical object will depend on its age, environment, and formation history. *See* ELEMENTS, COSMIC ABUNDANCE OF.

Stars and stellar evolution. The most numerous components of the nearby universe, as seen in the night sky, are the stars. These pinpricks of light are actually objects much like the Sun, but they appear faint due to their extreme distance from the Earth. Stars are enormous balls of hot gas (primarily hydrogen and helium), held together by their own gravitation. They are powered by nuclear reactions deep in their interiors, where temperatures and pressures are high enough to fuse hydrogen atoms together into helium, releasing energy in the process. (This process is sometimes referred to as hydrogen burning, although it is quite different from ordinary chemical burning.) A star's total mass, which may be from 0.08 to 100 times that of the Sun, determines nearly all of its other characteristics, including temperature, luminosity (intrinsic brightness), and how long it shines before running out of fuel. The Sun is an average star, with a surface temperature of about 5800 K, and a projected lifespan of 10^{10} years. More massive stars have higher core temperatures and pressures, so the fusion reactions take place much faster, and these stars shine brighter and hotter (thus appearing white or blue in color), expending their fuel reserves in only millions of years. Cool low-mass stars are more frugal with their available core hydrogen, and are expected to shine dimly but steadily, with a dull red color, for hundreds of billions of years. When plotted on a graph of temperature versus luminosity, this trend appears as a diagonal band known as the main sequence (**Fig. 1**). *See* CARBON-NITROGEN-OXYGEN CYCLES; MASS-LUMINOSITY RELATION; NUCLEAR FUSION; PROTON-PROTON CHAIN.

This main-sequence relationship holds only for stars which are burning hydrogen to helium in their cores. Once a star has used up all of its core hydrogen, it must change its internal structure to utilize new fuel sources. First, the hydrogen-burning zone expands outward from the core, and the enhanced power production causes the star to increase significantly in both luminosity and physical size, becoming a red giant or supergiant star. Further increases in core temperature and pressure permit the heavier elements to fuse and release energy, but these fuel sources are less efficient and relatively short-lived. A low-mass star (less than a few solar masses) will burn helium for a time, but as the helium fuel is exhausted and the internal furnace wanes, the outer layers of the star are ejected into space, and the core will gradually shrink into a tiny, dense ember, a white dwarf, glowing only from its residual heat.

Higher-mass stars will burn helium, then carbon, and then a succession of even heavier elements, each for a progressively shorter time, until fusable material of any sort abruptly runs out, and the star collapses catastrophically. Much of the interior mass of the star is compacted into an ultradense core; the outer layers rebound off this core and explode into space, forming a type II supernova, which shines for several weeks at a billion times the luminosity of the Sun. Many of the heavier chemical elements are formed in massive stars and during the supernova explosion itself, and are then spread throughout nearby space to be incorporated into future generations of stars. The stellar core is usually left behind as a neutron star, a small, rapidly rotating body consisting almost entirely of neutrons. Powerful magnetic fields on the surface of a neutron star can produce radio waves, which appear to blink on and off

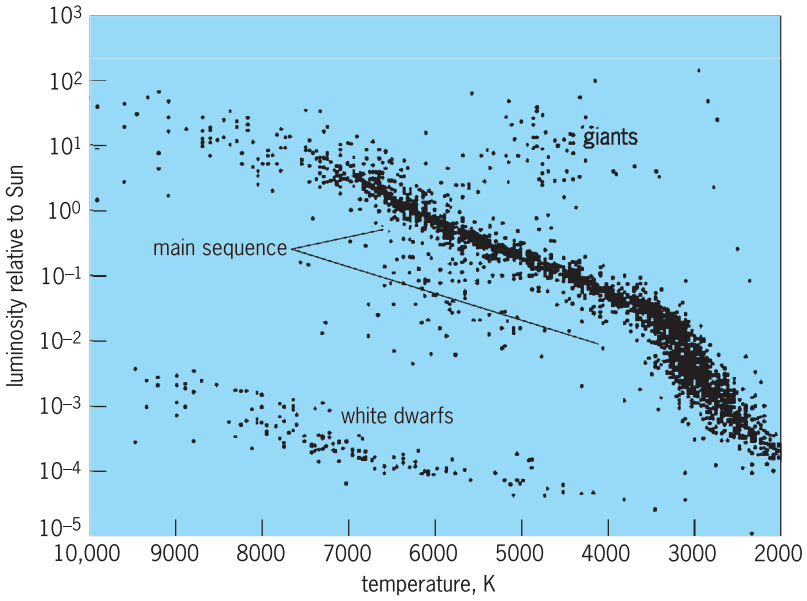


Fig. 1. Hertzsprung-Russell diagram, plotting absolute luminosity against temperature. Each dot represents a star in the local solar neighborhood. The diagonal band is the main sequence, where "normal" core hydrogen burning is taking place. Other regions of the diagram correspond to later stages of stellar evolution. (*Data from W. Gliese and H. Jahreiss, Preliminary Version of the Third Catalogue of Nearby Stars, Astron. Rechen-Institut, Heidelberg, 1991*)

as the star spins. A neutron star of this sort is called a pulsar. In the most extreme cases, the stellar core is compressed so far that it collapses to an infinitesimal point, forming a black hole, and the resulting release of energy is so great that the explosion can be seen from across the universe, appearing as a gamma-ray burst (GRB), a brief but exceptionally powerful pulse of high-energy radiation. The dense end states of stars—white dwarfs, neutron stars, and black holes—are collectively termed compact objects. *See* BLACK HOLE; GAMMA-RAY BURSTS; NEUTRON STAR; NUCLEOSYNTHESIS; PULSAR; STAR; STELLAR EVOLUTION; WHITE DWARF STAR.

Solar system. The Sun is accompanied by a number of smaller objects of various sizes and compositions. The Sun's gravitational domain extends out to almost half a parsec, but its planetary system lies much closer, within about 5×10^9 km (3×10^9 mi) of the center. The innermost planets—Mercury, Venus, Earth, and Mars—are small rocky bodies with solid surfaces and thin atmospheres. Slightly beyond the orbit of Mars is a belt containing several hundred thousand rocky asteroids, fragments of a planet that never quite formed. Farther from the Sun are the gas giants—Jupiter, Saturn, Uranus, and Neptune—with thick, turbulent, gaseous atmospheres surrounding small liquid or solid metallic cores. Pluto, usually considered the most distant of the "official" planets, is solid, but is composed predominantly of ice rather than rock, and is considered by some astronomers to be the largest and closest of a significant population of small icy bodies that orbit beyond Neptune in the Kuiper Belt. Over 750 of these Kuiper Belt Objects are currently known. Even farther outwards lies the Oort Cloud, a spherical shell containing yet more balls of ice, some of which occasionally plummet toward the inner solar system, get heated by the energy of the Sun, and appear as comets, as do some of the Kuiper Belt Objects. *See* ASTEROID; COMET; KUIPER BELT; PLANET; SOLAR SYSTEM.

Extrasolar planets. In recent years, it has become evident that the Sun is not the only star with a planetary system. Disks of particle matter have been seen around recently formed stars, and are thought to be analogous to the initial solar nebula out of which the Sun and the rest of the Earth's solar system formed. Moreover, over 100 nearby stars, most of them similar in type to the Sun, exhibit subtle periodic shifts in their motion, which are due to the gravitational influence of orbiting planetary companions. Most of these other solar systems are configured rather differently from the Sun's, with large planets orbiting close to their star, often even closer than Mercury is to the Sun. Other solar systems with smaller planets or large planets in slower, more distant orbits, may also exist, but measurement techniques must improve before such systems can be detected.

Interstellar material. The "empty" space between the stars actually contains significant amounts of matter—some of it distributed continuously, some of it concentrated in enormous dark clouds—collectively known as the interstellar medium. An atomic hydrogen component is distributed fairly smoothly, although with some variation in temperature and density. Also present are tiny solid dust particles, typically with diameters of 0.00025 cm (0.0001 in.) and smaller, and composed of graphite, silicon oxide, and complex carbon-chain and carbon-ring molecules. *See* INTERSTELLAR MATTER.

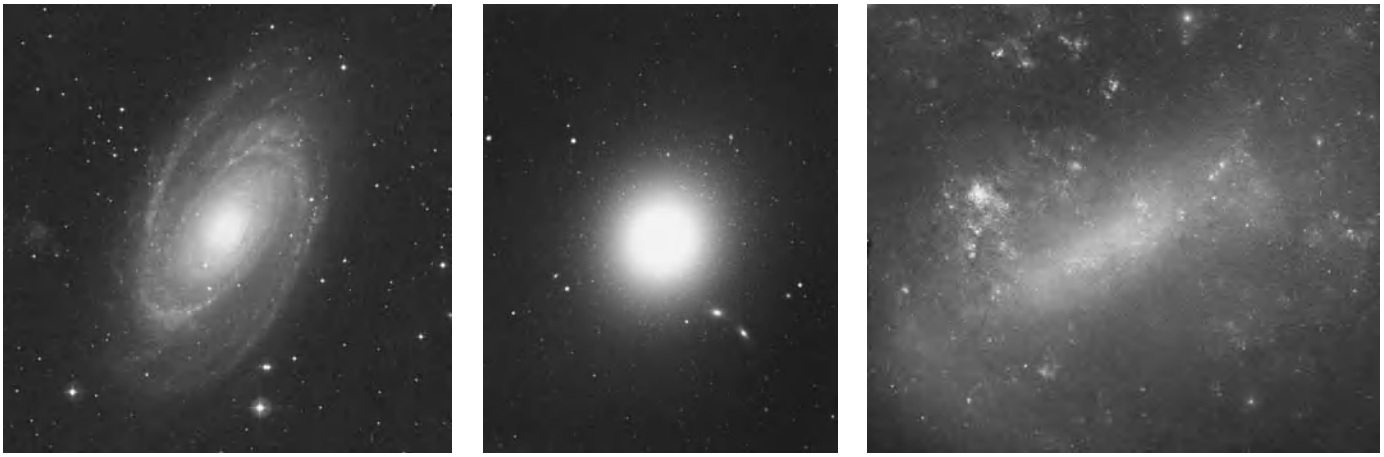
Also scattered throughout space are denser clouds of molecular hydrogen (H_2), carbon monoxide (CO), and more complex molecules, sprinkled with the heavier elements. Some of these clouds, or dark nebulae, are many parsecs in size, and contain enough raw material to form a million stars like the Sun. In fact, these clouds often are the site of star formation, when external gravitational disturbances or shock waves trigger the collapse of portions of the clouds. The densest clumps of mass within the cloud, having a stronger gravitational pull, will accumulate more of the surrounding material, increasing the pressure and temperature in the core of the clump to the point where hydrogen fusion can start. Star-forming regions, viewed with instruments such as the Hubble Space Telescope, are some of the most beautiful phenomena in the nearby universe, as new stars emerge from the residual gas surrounding their birthplaces. *See* MOLECULAR CLOUDS; PROTOSTAR.

Sometimes these clouds form self-gravitating spheres of gas which do not have the mass of approximately 0.08 solar mass that is required to support sustained internal fusion, and thus will not shine like stars, but which still glow dimly from the residual heat of their collapse. These "failed stars" or "superplanets" are called brown dwarfs, because they are very dim and red compared to real stars. A large number of brown dwarfs have been discovered since 1995, although not in sufficient numbers that they can account for the dark matter which appears to pervade the universe. *See* BROWN DWARF.

Interstellar space is also populated by high-velocity protons and electrons, which have been accelerated by supernova explosions and other mechanisms to nearly the speed of light. These energetic particles are called cosmic rays. A cosmic-ray proton, despite being just a single subatomic particle, can possess the equivalent momentum of a fast baseball. Most cosmic rays originate within the Milky Way, but some may have reached Earth from other galaxies. *See* COSMIC RAYS.

Galaxies. Stars and interstellar matter are not distributed uniformly throughout the universe, but cluster together in vast units known as galaxies, each containing from 10^7 to 10^{12} stars. Galaxies are categorized into three types—spirals, ellipticals, and irregulars—with numerous subclasses based on size and structure.

The Milky Way, the galaxy containing the Earth, is a typical spiral galaxy—a flattened disk of some 10^{11} stars (**Fig. 2***a*). An oblate central bulge, consisting mostly of older red stars, occupies the central region of the galaxy. The bulge is surrounded by a thin disk of younger stars. Within the disk, a pattern of spiral arms is defined by bright blue stars, recently formed from the gas clouds which also inhabit the disk. Both



(a) (b) (c)

Fig. 2. Galaxies. (*a*) Spiral galaxy M81 (NGC 3031). The flat disk, spiral arms, and central bulge are visible. The spherical halo that surrounds both is too sparsely populated to be evident (*Digital Sky Survey*). (*b*) Elliptical galaxy M87 (NGC 4486). The central galaxy appears quite spherical. Several hundred small globular clusters (each with thousands of stars) surround M87 (© *Anglo-Australian Observatory; photograph by David Malin*). (*c*) Large Magellanic Cloud, an irregular galaxy near the Milky Way. Bright regions containing hot young stars are found throughout it (© *Anglo-Australian Observatory; photograph by David Malin*).

the bulge and the disk are embedded in the halo, a sparsely populated sphere of very old stars, some of which date back to the formation of the galaxy. The Milky Way, like many other spirals, is accompanied by an assortment of dwarf galaxies and globular clusters (smaller collections of only thousands or millions of stars), which orbit the main galaxy. Dwarf galaxies which approach too closely to the galactic center are ripped apart by the galaxy's gravitational field, and their component stars are absorbed into the halo, suggesting that other such mergers in the past may explain the current size of the Milky Way. The visible portion of the disk extends about 30 kpc from center to edge, although the orbital motions of the stars and gas strongly imply a significant component of nonluminous dark matter both within the visible galaxy and beyond. Other spiral galaxies also exhibit this characteristic, implying that dark matter is a common constituent of galaxies. *See* MILKY WAY GALAXY; STAR CLUSTERS.

Elliptical galaxies are simpler systems than the spirals, consisting mostly of a singe population of cool low-mass stars, arrayed in a spherical or oblate configuration (Fig. 2*b*). Ellipticals have already converted all of their interstellar matter into stars, so that only the long-lived red stars now remain; young blue stars, like those found in spiral arms, have long since run out of fuel and expired. Ellipticals span a much wider range in size than the spirals—dwarf ellipticals, such as those orbiting nearby spiral galaxies, may contain only a few million stars, while the giant ellipticals that inhabit the centers of major clusters have upward of 10^{12} stars. As in spirals, the stars in ellipticals orbit the center of those galaxies at speeds which require significant additional dark mass, beyond that evident in the form of stars, to keep the galaxy from flying apart. Ellipticals are more likely than spirals to be found within galaxy clusters, which suggests that collisions or interactions between galaxies play some role in the formation of the elliptical type.

The irregular type includes all galaxies not readily recognizable as spirals or ellipticals (Fig. 2*c*). Irregular galaxies are similar to spirals in that they often exhibit active star formation, but they do not share the well-ordered structure of spirals. They may be the result of violent galaxy collisions, or perhaps are smaller galaxies that were prevented from developing structure due to gravitational disruption from a nearby larger galaxy. The Milky Way's two largest companions, the Large and Small Magellanic Clouds, are irregulars. *See* GALAXY, EXTERNAL; MAGELLANIC CLOUDS.

Quasars. Supermassive black holes, similar to the stellar-sized black holes which form during the supernova explosion of a massive star but containing millions or billions of solar masses instead of only a few, are thought to exist at the centers of many galaxies, including the Milky Way. Black holes themselves cannot, by definition, emit any light or other radiation, but they can serve as highly efficient engines for converting mass into energy. As stellar or interstellar matter is drawn into the black hole, it forms an accretion disk around the central point and is heated by friction to temperatures of millions of kelvins, so that it glows brightly with a luminosity equivalent to 10^{12} Suns. At the far reaches of the universe, these galaxy cores appear as faint starlike points of light, and were thus labeled quasistellar objects or quasars when first discovered. Nearby galaxies such as the Milky Way do not appear as quasars, because the flow of fuel to the black hole has slowed to a trickle, and the central engine shines only dimly. Farther away, however, the ultraluminous galaxy cores are seen as they were many billions of years ago, when the central engines were still being supplied with copious fuel. The most distant quasars known today are located more than 90% of the way to the edge of the observable universe. *See* QUASAR.

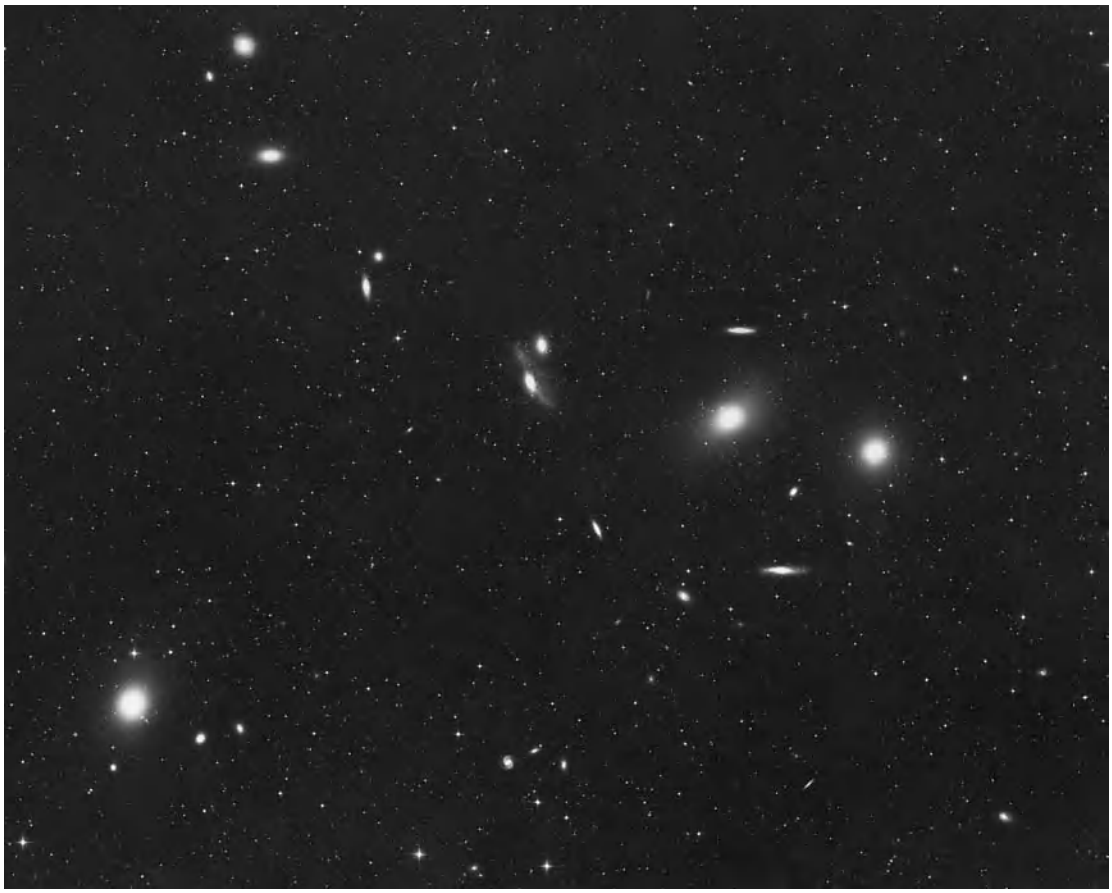


Fig. 3. Heart of the Virgo Cluster, the nearest moderately rich cluster of galaxies, including the giant elliptical galaxies M84, M86, and M87. (© *Anglo-Australian Observatory; photograph by David Malin*)

Groups, clusters, and large-scale structure. Galaxies themselves are usually bound together in groups (up to about 50 galaxies) and clusters (50 to thousands of galaxies) spanning regions 2–10 Mpc in diameter (**Fig. 3**). Like individual galaxies, groups and clusters are gravitationally self-contained systems, only on a larger scale. Clusters must contain more mass than is immediately evident—the total mass of the visible galaxies in a typical cluster is only a small fraction of the amount needed to keep the fast-moving galaxies from flying out of the cluster. There must therefore be some other mass component in addition to the galaxies, contributing its gravitational potential to keep the cluster together. Some of this "missing mass" can be attributed to hot diffuse gas, the intracluster medium, which is visible to x-ray and ultraviolet telescopes and fills the space between cluster galaxies, but the nature of the rest of the dark matter is yet to be explained. *See* LOCAL GROUP; VIRGO CLUSTER.

Clusters accumulate into yet larger entities called superclusters. Typical superclusters are on the order of 100 Mpc in size, and often take elongated or filamentary shapes. Recent redshift surveys, astronomical programs to measure the velocities of thousands or millions of distant galaxies and thus estimate their distance, have provided glimpses of these largest known levels of structure in the universe (**Fig. 4**). Groups and clusters seem to be concentrated in thin sheets, surrounding enormous voids with very few galaxies. Superclusters sit at the edges and vertices where surfaces intersect. The arrangement has been likened to a froth of soap bubbles, and may provide vital clues to the mechanisms by which structure formed out of the initially homogeneous early universe.

Antimatter. All subatomic particles have oppositely charged antiparticle counterparts. When a particle and its corresponding antiparticle collide, they annihilate, converting all their mass into energy. Both matter and antimatter are expected to have been formed in the early universe, but clearly not in precisely equal amounts, since the observable universe is predominantly matter. If slightly more matter were formed than antimatter, as predicted by current theory, then later matter-antimatter annihilation would leave behind only the "extra" matter, as observed today. Trace amounts of antimatter are still found in some astronomical environments, where new particle-antiparticle pairs are created out of free energy, in the opposite reaction from annihilation. Collisions between cosmic rays or high-energy photons can supply such energy, creating electrons and their antiparticles, positrons. *See* ANTIMATTER; ELECTRON-POSITRON PAIR PRODUCTION.

Nonbaryonic particles. Although most observable matter is baryonic in nature, several kinds of nonbaryonic matter also exist in the universe. Some

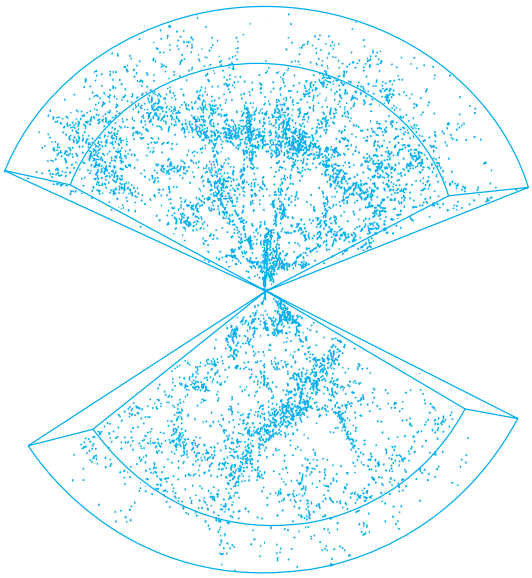


Fig. 4. Large-scale distribution of galaxies in two opposing regions of the sky, as measured by extensive redshift surveys. Each dot represents a galaxy. (*From L. N. Da Costa et al., A complete southern sky redshift survey, Astrophys. J., 424:L1–L4, 1994*)

nonbaryonic particles have been experimentally observed, and play a crucial role in explaining many processes, from nuclear fusion to supernova explosions. Other nonbaryonic particle types have been proposed on the basis of theory but have not yet been detected directly. According to current astronomical measurements, nonbaryonic matter accounts for approximately 23% of the total density of the universe, so it is important to determine the identity and characteristics of this component of the universe.

Electrons, strictly speaking, are not baryons, but they are considered as an integral part of ordinary atomic matter, and thus are usually not categorized with the more exotic nonbaryonic species described below. The net electric charge of the universe is assumed to be zero, so there are the same number of electrons as protons, but since the proton mass is almost 2000 times that of the electron, electrons make only a small contribution to the total cosmic mass. *See* ELECTRON.

A large number of other nonbaryonic particles, including mesons, muons, and taus, have been discovered in particle physics experiments. Many of these particles are quite massive (by subatomic standards), but they also have very short lifetimes—they rapidly decay into less massive particles. As such, they are considered unimportant for cosmology. *See* PARTICLE PHYSICS.

Neutrinos are electrically neutral, very low mass particles that are generated in nuclear reactions. They were once thought to be completely massless entities, but theoretical and experimental evidence now suggests that they have a tiny but nonzero mass, approximately 2×10^{-34} g (4×10^{-37} lb). Neutrinos do not interact strongly with baryons; they pass right through most ordinary matter, and even stars and planets are virtually transparent to them. This characteristic makes them useful tools for studying otherwise inaccessible environments, like the fusion core of the Sun or the centers of supernovae, but this also makes neutrinos difficult to detect and measure, as they usually pass through any sort of detection device without leaving a trace. Projects to detect neutrinos from the Sun and other astronomical sources have met with some success in recent years, offering an entirely new means of studying the universe. *See* NEUTRINO; NEUTRINO ASTRONOMY; SOLAR NEUTRINOS.

According to new theories of particle physics, a variety of other nonbaryonic particle species might also exist. Some of these hypothetical particles could be much heavier and slower than neutrinos, but like neutrinos, would only interact with baryonic matter via the gravitational force. The most likely candidate to fill this role is the so-called weakly interacting massive particle (WIMP). Lightweight axions are another possibility, along with magnetic monopoles, WIMPzillas, and neutralinos. Experiments attempting to directly detect WIMPs and other particles are currently under way, but have not yet proven the existence of any of these exotic entities. *See* MAGNETIC MONOPOLES; WEAKLY INTERACTING MASSIVE PARTICLE (WIMP).

Dark matter. A large fraction of the mass of the universe is in some form which cannot be seen, but which is evident from its gravitational effect on the motions of bright objects, such as stars and galaxies. This dark matter, or "missing mass," is present on many different scales, from galaxies to the universe as a whole. Several possibilities for the nature of the dark matter have been suggested. Numerous small dim baryonic objects, such as faint red stars, brown dwarfs, or black holes, located in the halos of galaxies, might account for some of the galactic-scale dark matter. However, searches for these MACHOs (massive compact halo objects) have turned up far fewer confirmed detections than necessary to explain all the missing mass. Furthermore, several independent lines of argument indicate that baryons can comprise only about 4% of the total mass-energy density of the universe, while the dark matter has to make up approximately 23% of the cosmic total in order to properly explain the distribution and motion of galaxies. Dark matter must therefore consist primarily of nonbaryonic particles. Neutrinos were once considered as likely candidates, but recent observations of the cosmic microwave background argue against this possibility, because neutrinos travel at nearly the speed of light and would spread out too quickly to permit the observed lumpiness of matter in the early universe. Slower-moving ("cold") particles such as WIMPs are therefore favored by current theories. Active theoretical and experimental efforts are under way to determine the characteristics of such particles, and the effect they would have on the distribution and dynamics of the luminous matter, but direct detection of these particle species will be very challenging. The identification of dark matter remains one of the key unresolved issues of modern astrophysics. *See* DARK MATTER.

Energy. Energy is a physical entity as real as matter, but somewhat less tangible, which makes it more difficult to categorize easily. Like matter, energy comes in many different forms, which can be readily transformed from one to another. Energy can also be converted to and from matter. Various forms of energy which play a significant role in the universe—gravitational, electromagnetic, kinetic, nuclear, and dark energy—are described below. *See* ENERGY.

Gravity is the predominant physical force in most astronomical contexts, and is thus responsible for the existence of most of the universe's structure, including stars, planets, galaxies, clusters, and superclusters. Gravitational potential also drives many of the key energy-generating mechanisms in the cosmos, such as nuclear fusion in stellar cores and accretion disks in quasars. *See* GRAVITATION.

Energy is readily transmitted throughout the universe via electromagnetic radiation, which may take the form of visible light, radio waves, x-rays, gamma rays, or any other part of the electromagnetic spectrum. Electromagnetic energy may be described either as waves—intertwined electric and magnetic fields rippling through empty space—or as particles called photons, which travel in straight lines like small solid objects. The universe now contains about 400 photons per cubic centimeter (6500 per cubic inch), so photons outnumber atoms by a large margin. *See* ELECTROMAGNETIC RADIATION; LIGHT; PHOTON.

Matter may contain energy (in addition to the rest-mass energy defined by relativity) by virtue of its motion. This kinetic energy may be due to large-scale bulk motion, such as a planet hurtling through space, or it may be in the form of thermal energy—the microscopic motion of an object's constituent atoms, vibrating or moving relative to each other. The higher the temperature, the faster the average motion, and the more energy is present. *See* KINETIC THEORY OF MATTER.

Large reserves of energy are stored in the force which holds atomic nuclei together. There are actually two such forces, the strong nuclear force and the weak nuclear force, which play different roles at the subatomic level. Although considerably weaker than the strong force, the weak nuclear force plays an important role in radioactive decay and the formation of isotopes of the heavy elements. The strong force maintains the nuclear structure upon which matter is based, but some of this energy can be released in other forms during nuclear reactions. When two hydrogen atoms are fused together to form helium, for instance, the hydrogen nuclei have a greater total binding energy than the resulting helium nucleus, and the difference is released as free energy (in the form of photons and neutrinos). Nuclear fusion is the process which powers all luminous stars in the universe. Energy release can also take place when a large heavy nucleus splits into two lighter nuclei with less net energy. This is the mechanism behind nuclear fission, as found in nuclear power plants and uranium/plutonium bombs. But because of the relative scarcity of heavy fissionable elements in the universe, fission does not play any appreciable role in astronomical environments. *See* NUCLEAR FISSION; NUCLEAR REACTION; RADIOACTIVITY; STRONG NUCLEAR INTERACTIONS; WEAK NUCLEAR INTERACTIONS.

Research suggests that, on cosmological scales, gravitational attraction is countered by a previously undetected form of energy, known as dark energy or the cosmological constant, which causes space to expand. A force of this sort had originally been suggested by Einstein's equations of general relativity, but the strength of this force was assumed to be zero, because no excess expansion rate had been observed in the motions of galaxies. Since 1998, however, observations of distant supernovae and the cosmic background radiation have led astronomers to the conclusion that the expansion of the universe is indeed accelerating, and that dark energy is, in fact, the dominant component of the universe. According to these data, the dark energy comprises 73% of the total density of mass and energy in the universe, "outweighing" the dark matter and baryonic matter by a factor of 3. The actual nature of the dark energy is still a mystery. *See* COSMOLOGICAL CONSTANT; DARK ENERGY; RELATIVITY.

The traditional four basic forces of nature—gravity, electromagnetism, and the strong and weak nuclear forces—may in fact all be different facets of a single fundamental force. Particle physics experiments show that under conditions of extremely high energy the weak nuclear force and electromagnetism merge into a single "electroweak" force. Theoretical efforts are being made to devise a grand unified theory under which the strong force is also incorporated, and a still more general force that also includes gravity. Such a unified force may have existed during the moments following the big bang, and then fragmented into separate forces as the universe expanded and cooled. With the discovery of the long-range "antigravity" effect of dark energy, the task of combining all of the known physical forces into a single coherent framework has become even more challenging. *See* ELECTROWEAK INTERACTION; FUNDAMENTAL INTERACTIONS; GRAND UNIFICATION THEORIES; STANDARD MODEL; SUPERSYMMETRY.

Origin, Evolution, and Fate

The universe is a dynamically evolving system; there are compelling reasons to believe that it has not always been in its current state. Most of the evidence for this assertion is indirect, as humans have not observed the cosmos for long enough to actually see things change on large scales. But by closely studying the current distribution and motion of matter and energy, and collecting the "fossil light" from distant objects, scientists have constructed a consistent picture of the creation of the universe, the big bang theory, which explains the observations fairly well. Many of the specifics are still unresolved, however, and are currently the subject of research and debate.

Big bang. The observable universe originated 13.7 billion years ago in a fiery cataclysm termed the big bang. This was not an explosion of compressed

matter and energy into a previously empty space. Instead, space itself, as well as everything contained therein, sprang from a single point of near-infinite density and temperature, and grew to the volume observed today. As it expanded, it cooled, allowing familiar forms of matter to condense from the high-energy "soup" of energy and subatomic particles that constituted the very early universe. The framework for the present-day distribution and dynamics of matter in the universe was also established at these very early times, and thus remnants of the extreme conditions that existed at the beginning can still be seen. Much of the current understanding of the big bang and the early universe hinges on two key observations: velocities of distant galaxies and the cosmic background radiation. *See* BIG BANG THEORY.

Cosmological redshifts. The relative velocity of an astronomical object, toward or away from an observer, can be determined very precisely by measuring the color distribution of the light from the object. Distinct spectral features, characteristic of the different chemical elements that are present, appear at well-established locations in the spectrum, but the entire spectrum is shifted toward shorter wavelengths (bluewards) for an object approaching the observer, or toward longer wavelengths (redwards) for a receding object. The amount of this Doppler shift is closely related to the actual relative speed of the object. *See* DOPPLER EFFECT.

If galaxies were moving randomly in the universe, observers would expect to see just as many blueshifted galaxies as redshifted galaxies, roughly half moving toward Earth and half moving away. But measurements of galaxy motions made in the early twentieth century showed a very different distribution: With the exception of a few nearby galaxies, all galaxies exhibited redshifts of varying amounts, implying that they were all moving away from Earth. Edwin Hubble extended this work by determining the distances to these galaxies, and found that recession velocity was directly proportional to distance—that the more distant galaxies were moving away faster (**Fig. 5**). The constant of proportionality relating speed and distance now bears Hubble's name, and an accurate determination of this Hubble constant (H_0) has been one of the central pursuits of modern astronomy for the past several decades. Recent studies converged to a value of $H_0 = 72$ kilometers per second per megaparsec.

The observation that nearly all galaxies are moving away from Earth's location in space might suggest that Earth is at a special location in the universe, at or near the center of an expanding distribution of matter. It is highly unlikely, however, that Earth would happen to be located at such a special position. Fortunately, these observations can be explained in another fashion. If space itself is expanding uniformly in all directions, and the galaxies are being carried along in this general expansion, then any point in the universe would see all other points moving away. The greater the distance between two points, the more space exists between them, and the faster this distance increases. Hubble's law is therefore a consequence of the uniform expansion of space, which causes more distant galaxies to exhibit larger redshifts because they are receding from the observer faster.

If space is expanding uniformly in all directions, then in the past the universe was smaller. All the matter and energy that is observed in the current-day cosmos would be present in this smaller volume, so the density (average mass per volume) and temperature (related to the average energy per volume) of the universe would be higher than they are today. At even earlier times, the volume would be yet smaller, and the density and temperature even higher, and at some point in the past all matter and energy may have existed in a single point of infinite temperature and density. *See* HUBBLE CONSTANT; REDSHIFT.

Cosmic background radiation. If this picture is correct, and the current universe was spawned from just such a primordial fireball, then the sky should be filled with a residual afterglow from the era when the matter in the universe was hot and emitting strongly, in much the same way that the surfaces of stars shine today. This primordial radiation will have cooled dramatically, though, since the space containing it has expanded since the radiation was emitted. From its initial characteristic temperature of about 3000 K, this cosmic background radiation, or cosmic microwave background, is expected to have cooled to about 3 K above absolute zero, so that instead of existing mostly as visible light, the energy distribution will have shifted far to the red, into the microwave radio region of the spectrum.

This microwave background was first detected by A. Penzias and R. Wilson in 1965 as persistent low-intensity static in microwave communications equipment, and was identified as being of cosmic origin. Subsequent ground-based measurements of the energy distribution of the cosmic background radiation suggested a characteristic temperature of approximately 3 K, but it was not until the *Cosmic Background Explorer* (*COBE*) satellite mission

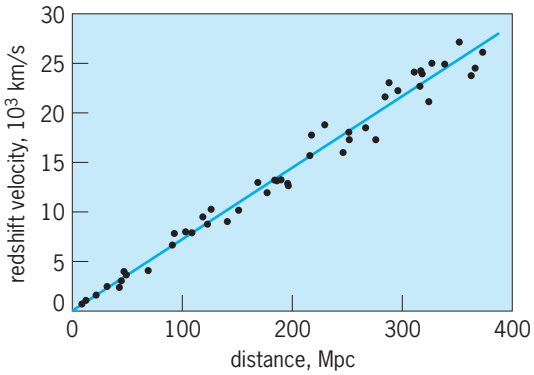


Fig. 5. Hubble diagram showing linear relationship between galaxies' distances and recession velocities. Each point represents one galaxy. The slope of the best-fit line gives the Hubble constant H_0. (*Data from A. G. Reiss et al., Obsevational evidence from supernovae for an accelerating universe and a cosmological constant, Astron. J., 116:1009–1038, 1998; and S. Perlmutter et al., Measurements of Ω and Λ from 42 high-redshift supernovae, Astrophys. J., 517:565–587, 1999*)

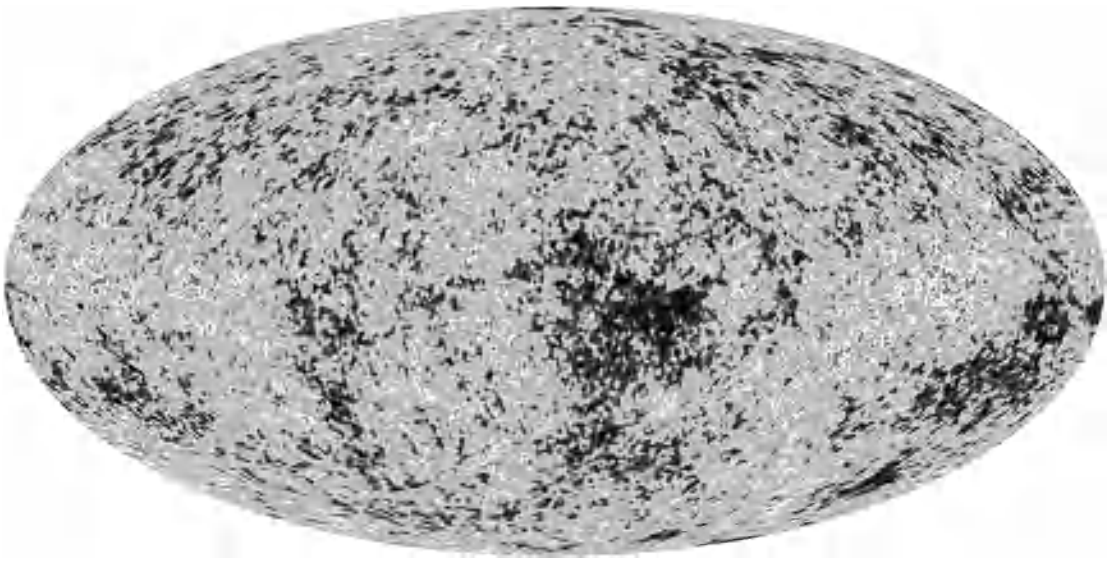


Fig. 6. Full-sky projection map of temperature variations in the cosmic background radiation, as measured by the *Wilkinson Microwave Anisotropy Probe* (*WMAP*). (*NASA/WMAP Science team*)

in 1990 that the spectrum was measured simultaneously over a wide range of wavelengths, firmly establishing the cosmic background radiation temperature at 2.726 K.

Since the cosmic background radiation is an intrinsic characteristic of the universe, observers expect to see it reaching the Earth uniformly from all directions in space, and the *COBE* measurements confirmed that this is indeed the case, once the Doppler effect from the Earth's motion through space is subtracted. Subsequent observations of the cosmic background radiation by highly sensitive microwave telescopes, including the *Wilkinson Microwave Anisotropy Probe* (*WMAP*), have revealed faint ripples in the temperature distribution of the background radiation. These patterns in the cosmic background radiation are a result of the very first clumps of matter that accreted in the early universe, as regions with slightly different densities emitted slightly hotter or cooler radiation. By measuring the amplitude and angular size of these temperature variations, researchers have been able to deduce the age, total density, and early history of the universe with unprecedented precision. *See* COSMIC BACKGROUND RADIATION; WILKINSON MICROWAVE ANISOTROPY PROBE.

Evolution of the universe. The best models for the beginning of the universe start at 10^{-43} s after the big bang itself. The conditions before this point were so extreme that the current understanding of physics is insufficient to say anything meaningful (although new theories are being developed to address this limitation). At 10^{-42} s, the universe was 10^{45} times smaller than it is today (approximately 6×10^{-18} cm, or 2×10^{-18} in.), and had a mean temperature of 10^{30} K. At such temperatures, matter as presently known cannot exist, because the energies are so enormous that even the protons and neutrons themselves are torn apart into separate sub-subatomic particles called quarks. The universe was a featureless mixture of subatomic particles and high-energy photons. *See* QUARKS.

As the universe expanded and cooled, however, more particles were permitted to form, once they could exist without being immediately destroyed. At 10^{-34} s, when the temperature had dropped to 3×10^{26} K, heavier exotic particles like magnetic monopoles could have emerged. Around this time, the rapidly enlarging structure of space may have undergone an era of even faster expansion, driven by the energy of space itself. This era of hyperfast inflation helps to explain features of the present-day observable universe, such as the uniformity of the cosmic background radiation over the entire sky, and the way in which the average mass density of the universe is high enough for structures such as stars and galaxies to form, but not so large that it would immediately recollapse upon itself. *See* INFLATIONARY UNIVERSE COSMOLOGY.

When the temperature had dropped further, to about 3×10^{12} K, protons and neutrons condensed out of the quark mixture. It was still much too hot for electrons to join them and form atoms, but more complex atomic nuclei—deuterium (proton plus a neutron), helium-3 (2 protons plus a neutron), helium-4 (2 protons plus 2 neutrons), and lithium (3 protons plus 4 neutrons)—were created, in a fashion similar to stellar core fusion. This era of big bang nucleosynthesis established the initial composition of the universe, about three-quarters hydrogen, one-quarter helium, and a smattering of lithium, from which all subsequent stellar and supernova nucleosynthesis has proceeded.

Temperatures were still too high for electrons to join these nuclei to make complete atoms, until a time about 370,000 years after the big bang. Up to this point, the universe was relatively opaque, since unattached electrons are very good at absorbing light and other electromagnetic radiation. Once the temperature fell below 3000 K, however, protons

and electrons could combine into neutral hydrogen atoms, and the universe suddenly became transparent to most wavelengths. The cosmic background radiation formed at this point, and has had little interaction with matter since, which is why observations with microwave telescopes such as *COBE* and *WMAP* reveal such detailed information about this epoch.

The formation of structure—the development of clumpiness in the universe, which appears today as stars, galaxies, and clusters—started slightly prior to matter-radiation decoupling. These very first clumps of mass left their faint imprint on the temperature distribution of the cosmic background radiation, and then continued to accumulate and grow after becoming independent from the radiation field. The exact sequence of structure formation—whether many smaller objects such as stars formed first and then accumulated to form galaxies and clusters, or whether matter clumped on galaxy- or cluster-sized scales and then subdivided into smaller bodies—is still not well understood, because of the scarcity of luminous objects and the difficulty of detecting them at such extreme distances. Sophisticated computer simulations, combined with indirect evidence from observations of the cosmic background radiation and the most distant quasars, suggest that massive stars (100 times the mass of the Sun, or greater) formed relatively early and then exploded as supernovae, pumping energy and newly formed heavier elements into the surrounding clouds of hydrogen and helium, and thus setting the stage for the formation of galaxies and stars observed today. Observations from future generations of space telescopes will help resolve many of the uncertainties regarding this era of cosmic history.

Eventually, within 10^9 years of the big bang, larger structures such as galaxies did take shape, and within them subsequent generations of stars coalesced from interstellar clouds of hydrogen. The Sun (along with its system of planets) appeared about 5×10^9 years ago. In elliptical galaxies, nearly all the gas has been turned into stars by now, while spiral and irregular galaxies continue to spawn new stars at a steady rate. Large galaxies gradually absorb and consume any dwarf galaxy companions that approach too closely, and occasionally two large galaxies will collide and merge. Meanwhile, the overall expansion of the universe continues.

Ultimate fate of the universe. After another 5×10^9 years, the Sun will run out of hydrogen fuel in its core, and after swelling up into a red giant, will collapse into a white dwarf, gradually fading in brightness as it cools. Most of the other stars that are seen in the night sky will meet a similar fate, some of them within a few million years, others lasting hundreds of billions of years before their fuel reserves run out. New generations of stars will be formed to take their place, but over time the interstellar hydrogen reserves of spiral galaxies such as the Milky Way will be exhausted, and the formation of new stars will cease, leaving only the long-lived dim red stars.

On larger scales, the universe will probably continue to expand indefinitely. It was once thought that if the average cosmic density of matter were high enough, the mutual gravitational attraction would be sufficiently strong to gradually slow the expansion to a standstill, and then cause the universe to contract, eventually collapsing in a "Big Crunch" (or "Gnab Gib"). According to the current tally of baryonic and dark matter, however, there is not enough gravitational mass in the universe to overcome the current rate of expansion. With the added repulsive force of the dark energy, the expansion rate will increase, instead of decrease. Objects such as stars, galaxies, and clusters, which are kept together by their own self-gravity, will remain largely intact, but the enormous voids separating clusters and superclusters from each other will grow ever larger. Eventually, over time scales of trillions of years, the universe will grow empty and dim, illuminated by only a few last stellar remnants, cooling slowly toward absolute zero.

Conclusion. Over the past several centuries, astronomers have assembled a fairly comprehensive and consistent view of the universe. Many of the major constituents are now known, and modern astronomical theories offer plausible explanations of how they were formed, what they are made of, and how they interact with each other. There are, however, a number of key unanswered questions, such as the nature of dark matter and dark energy, which will require many further advances in technology, observation, and theory to fully resolve. *See* COSMOLOGY.

Bradford B. Behr

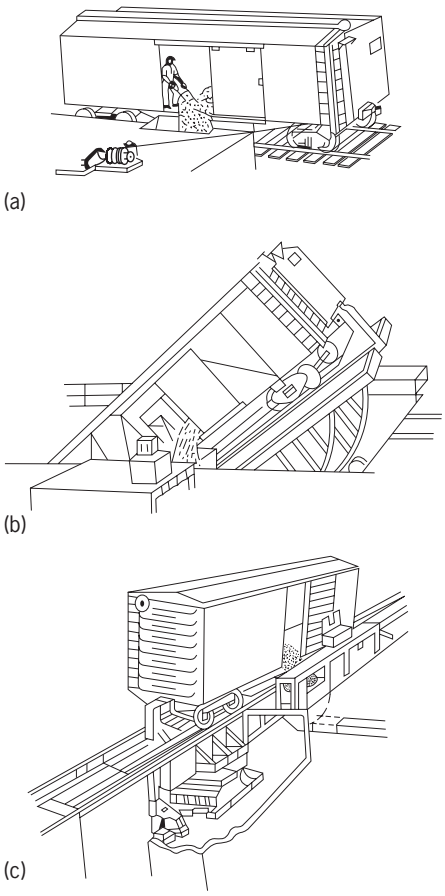
Bibliography. B. W. Carroll and D. A. Ostlie, *An Introduction to Modern Astrophysics*, Addison-Wesley, 1996; E. W. Kolb and M. S. Turner, *The Early Universe*, Addison-Wesley, 1990.

Unloader

A power device for removing bulk materials from railway freight cars or highway trucks. Especially in the case of railway cars, the car structure may aid the unloading, as in the hopper-bottom dumper; here the unloader may be a vibrator which improves flow of the bulk material into a storage bin or to a conveyor. Thus an unloader is the transitional device between interplant transportation means and intraplant materials-handling facilities. *See* MATERIALS-HANDLING EQUIPMENT; TRANSPORTATION ENGINEERING.

Basically, design considerations for an unloader are similar to those for a device for removing bulk from a storage bin or warehouse with low headroom and limited access. A special consideration is the possible need for an elevating platform to bring the car or truck floor to the level of the unloading dock and to maintain the alignment as material is removed and the vehicular springs expand. Unloading may then be accomplished by portable belt, drag-chain, flight, or similar conveyor (**illus.** *a*). *See* BULK-HANDLING MACHINES.

A fully mechanized unloader includes a section of track to which the car is clamped and then



Unloader. (*a*) Power scoop. (*b*) Rocker. (*c*) Tilter. (*Link-Belt Co.*)

rocked or tilted (illus. *b* and *c*). In this way unloading rates of up to ten boxcars per hour are achieved. For less specialized unloading, dock equipment such as power scoops may be used, or yard handling equipment such as boom unloaders may be adapted.

Frank H. Rockett

Upper-atmosphere dynamics

The motion of the atmosphere above 50 km (30 mi). The predominant dynamical phenomena of the upper atmosphere are quite different from those encountered in the lower atmosphere. Among those encountered in the lower atmosphere are cyclones, anticyclones, tropical hurricanes, thunderstorms and shower clouds, tornadoes, and dust devils. Even the largest of these phenomena do not penetrate far into the upper atmosphere. Above an altitude of about 50 km (30 mi), the predominant dynamical phenomena are internal gravity waves, tides, sound waves (including infrasonic), turbulence, and large-scale circulation. These are described below.

Internal gravity waves. Except under meteorological conditions characterized by convection, the atmosphere is stable against small vertical displacements of small air parcels; this results from buoyancy forces that tend to restore displaced air parcels to their original levels. An air parcel therefore tends to oscillate around its undisturbed position at a frequency known as the Brunt-Vaisala frequency; the square of this frequency is given by Eq. (1), where

$$\omega_B^2 = -g\left(\frac{\rho_0'}{\rho_0} + \frac{g}{C^2}\right) \tag{1}$$

ρ_0 is the undisturbed density, ρ'_0 is the vertical gradient of the density, g is the acceleration of gravity, and $C = (\gamma p_0/\rho_0)^{1/2}$ is the speed of sound; γ is the ratio of the specific heats, and p_0 is the undisturbed pressure. Typical periods for the Brunt-Vaisala oscillation are in the vicinity of 5 min.

If pressure waves are generated in the atmosphere with frequencies much greater than ω_B, they propagate as sound waves. For frequencies much less than ω_B, the waves propagate as internal gravity waves; in this case, the restoring forces for the wave motion are provided primarily by buoyancy (that is, gravity) rather than by compression. These waves have peculiar properties. The wave motion is nearly horizontal, but the inclination from the horizontal is sufficient so that gravity provides the restoring force for the displaced air parcels. The phase velocity of the waves is downward, but the group velocity is upward. Such waves are presumably excited near the Earth's surface, probably in wind systems, and the energy flow is upward, reaching well into the ionosphere. Owing to the decrease in density with altitude, the amplitudes of the waves must increase with altitude in order to preserve continuity of energy flow; above 60 km (36 mi), the amplitude is so large that it dominates the observed wind profiles. An example is given in **Fig. 1**, in which a wavelike structure with a vertical wavelength of about 20 km (12 mi) can be seen. In the lower atmosphere, the gravity waves are presumably also present, but with amplitudes so small that they cannot be detected.

Above 100 km (60 mi), dissipative mechanisms come into play for internal gravity waves; these include viscosity, thermal conductivity, and eddy transport of heat and momentum. The dissipation of the waves, judged from their observed amplitudes below 100 km (60 mi), provides an energy source which is apparently large enough to be of some significance in the heat budget of the upper atmosphere.

Internal gravity waves should not be confused with surface waves, which also occur within the atmosphere. The surface waves occur at discontinuity surfaces, or in regions where the gradient is strong enough to guide the propagation of the wave in a direction perpendicular to the gradient. Such waves are frequently visible on cloud sheets.

Tides. Tides are internal gravity waves of particular frequencies. The term tidal usually implies that the exciting force is gravitational attraction by the Moon or Sun. However, it is conventional in the case of atmospheric tides to include also those waves that are excited by solar heating. One is therefore concerned with three separate excitation functions—lunar gravitation, solar gravitation, and solar heating.

Examination of barometric pressures in tropical latitudes normally discloses two maxima and two

minima per day, the maxima occurring near 1000–2200 h local time; at higher latitudes the variations are smaller and the noise due to passing weather systems is much larger, and statistical analyses are needed to find the tidal oscillations. The gravitational tidal force of the Moon is 2.2 times larger than that of the Sun, and its period is 12 h 26 min. The fact that only the 12-h oscillation is readily seen on a barograph record therefore indicates that the oscillation is of thermal origin, or that the oscillation is so sharply resonant that the 12-h tidal force of the Sun excites a much stronger oscillation than the larger 12-h 26-min tidal force of the Moon. The possibility of a thermally excited oscillation of period 12 h arises because the solar heating curve has harmonics, that is, it is not a simple sinusoidally oscillating function. Although the possibility of a resonant excitation of the 12-h oscillation by solar gravitational force has been considered and the resonant period of the atmosphere is about 12 h, the resonance is now thought to be much too broad to discriminate between the solar and lunar gravitational tidal periods. Both are amplified several times by the resonance, but the thermal excitation must be responsible for the major observed effect.

Although the lunar tide of period 12 h 26 min is not discernible on a barograph trace, it can be found from a statistical analysis, and it is about 15 times smaller than the 12-h oscillation.

Although tidal oscillations are small at the Earth's surface, they are relatively large at high atmospheric levels. Only a few days' observations of the winds near the altitude of 100 km (60 mi) is required to disclose the diurnal pattern of behavior in which the largest components have 24- and 12-h periods; the amplitudes of these tidal winds are about 100 times greater than those at the Earth's surface. The tidal wind patterns in the upper atmosphere also disclose their presence in another way, by generating electrical currents in the ionosphere through a dynamo action. These in turn give rise to diurnal variations in the geomagnetic field that can be observed at the Earth's surface. The lunar tidal oscillations also increase with altitude, but they are smaller than the solar tidal oscillations (about 2 m/s or 4 knots as compared to 20 m/s or 40 knots), so that they have not been recognized in the limited amount of wind data available. However, they are recognizable in terms of the geomagnetic variations that they produce, although a statistical analysis is needed to find them. *See* ATMOSPHERIC TIDES; GEOMAGNETIC VARIATIONS; IONOSPHERE.

Sound waves and infrasonics. Sound waves generated in the lower atmosphere may propagate upward; to maintain continuity of energy flow, the waves might be expected to grow in relative amplitude as they move into the more rarefied upper atmosphere. However, higher temperatures in the upper atmosphere refract most of the energy back toward the Earth's surface, giving rise to the phenomenon known as anomalous propagation. This involves the redirection of upward-moving sound waves back to the surface beyond the point where the source (usually a large explosion) can be heard by waves propagating along the surface. Thus there is normally a zone of silence, beyond which the source can again be heard because of anomalous propagation.

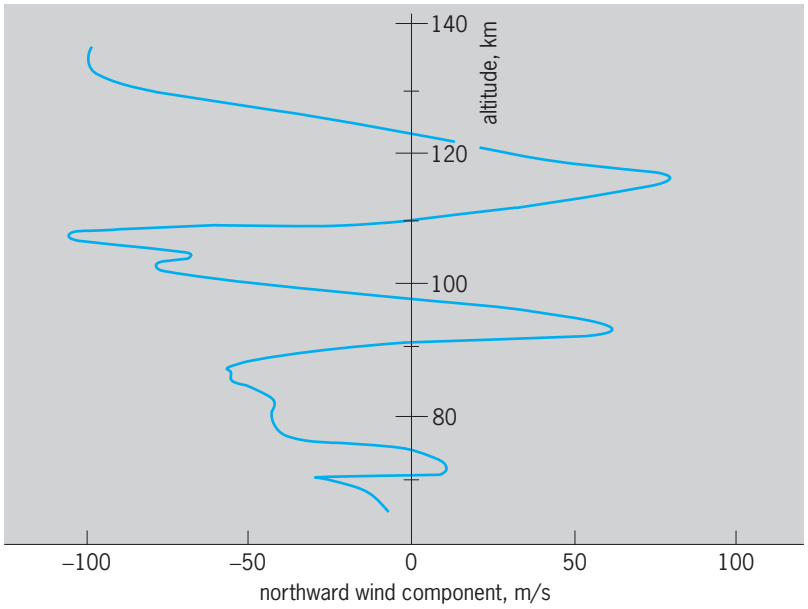


Fig. 1. Northward wind component over Wallops Island, Virginia, on June 6, 1962. 1 km = 0.6 mi; 1 m/s = 2 knots.

Sound waves or infrasonic waves (frequencies too low to be audible) generated by a thunderstorms have been considered possible sources of energy input in the upper atmosphere as the waves become shock waves on entering more rarefied regions of the atmosphere. The refraction effect mentioned above makes this unlikely, but the possibility has remained under consideration. The energy spectrum of sound waves generated by thunderstorms has been measured. A broad peak near 200 Hz was found, with no energy increase in the infrasonic region.

Infrasonic waves with periods from 20 to 80 s have been observed occasionally with detectors at the Earth's surface in connection with auroral activity, the average rate of observed occurrence being about once a day. The occurrence seems to be associated with rapidly moving auroral forms where the rate of movement is supersonic. Although this rapid motion does not represent a supersonic motion of atmospheric gas, but rather only a movement of the energetic particle source causing the aurora, nevertheless, a shock wave should be produced by the supersonic motion of the generating mechanism, the aurora. The observed direction of arrival agrees with this concept of an auroral source, the preferred direction of propagation being the direction of motion of the auroral forms. The observed amplitudes range up to 10 dynes/cm^2 (1 pascal). *See* SOUND.

Turbulence. There is clear visual evidence of turbulence in the upper atmosphere; this evidence is obtained by examination of vapor trails released from rockets or of long persisting meteor trails. The trails spread above about 110 km (66 mi) by an amount

that corresponds to molecular diffusion, but at lower altitudes the spreading is far in excess of that which can be attributed to this mechanism, and turbulence is the only known means of explaining the additional spreading. Further, in the region below 110 km (66 mi), the trails adopt a turbulent appearance, whereas at higher altitudes, where molecular diffusion predominates, the trails are smooth, as should be expected.

The source of the turbulence is not clear. The atmosphere is thermodynamically stable against vertical displacements throughout the region above the troposphere, and work has to be done against buoyancy forces in order to produce and maintain turbulence. The only apparent source of energy is internal gravity waves, either tidal or of random period. One concept is that wind shear associated with such waves may become great enough to overcome the stability and produce turbulence; the relevant criterion for this is that the Richardson number be smaller than some arbitrary value. The Richardson number essentially expresses the ratio of the stability (the work that must be done against buoyancy forces if turbulence is to be maintained) to the destabilizing forces involved in wind shear. It appears probable that the criterion is satisfied occasionally, but the trails do not give the impression of being turbulent only in the regions of strong shear. Another concept is that gravity waves may at times be of sufficiently large amplitude so that potentially colder air is brought over potentially warmer air; that is, the temperature distribution may be distorted in such a way that the adiabatic lapse rate is exceeded in some regions so that thermal convection can proceed. Again, the trails do not have the appearance of being convectively unstable in narrow regions as the concept requires.

The presence of turbulence and its associated eddy transport cannot be doubted, even though there are difficulties in understanding the energy source and the physical generation mechanism. In addition to the visual evidence provided by vapor and meteor trails, the effect of the eddy transport on atmospheric structure is such that much can be deduced about the average rate of eddy transport from a study of atmospheric structure. On this basis it is recognized that the average eddy diffusion coefficient varies from about 10^4 cm^2/s just above the tropopause up to about 10^7 cm^2/s near 110 km (66 mi), above which it is exceeded by the molecular diffusion coefficient (above 110 km, the eddy coefficient probably decreases because of the severe damping of eddies by viscosity and heat conduction). *See* DIFFUSION.

The use of an eddy diffusion coefficient permits an evaluation of the transport properties of eddies without specifying their physical properties. The generally accepted concept is that the larger eddies break up into progressively smaller eddies, which finally dissipate their energy against viscous forces. The largest eddies are responsible for most of the transport. Examination of vapor trails suggests that the largest eddy size is of the order of 100 m (330 ft); with eddies of this size, eddy velocities near 10 m/s (20 knots) would be required to produce the observed mixing rates, and the lifetimes of the largest eddies would be a few tens of seconds. *See* METEOROLOGICAL ROCKET.

Large-scale circulation. There are prevailing patterns of atmospheric circulation in the upper atmosphere, but they are very different from those that occur in the lower atmosphere, which are associated with weather systems and have complicated structures resulting from growth of instabilities. The upper atmospheric large-scale wind systems are mainly diurnal in nature and global in scale.

The main heat source that is responsible for the upper atmospheric circulation is ultraviolet radiation from the Sun, radiation that is mainly absorbed at altitudes between 100 and 200 km (60 and 120 mi). The atmosphere is not a good infrared radiator in this altitude region, so the temperature rises rapidly with altitude, providing a temperature gradient of such a magnitude that molecular conduction transfers the absorbed heat downward to altitudes below 100 km (60 mi) where the atmosphere does have the capability of radiating the energy back to space. **Figure 2** shows some typical temperature distributions through the atmosphere. Above about 300 km (180 mi), the temperature becomes roughly constant with altitude because very little energy is absorbed there (the gas is exceedingly rarefied) and the thermal conductivity is good enough under these circumstances to virtually eliminate vertical temperature gradients. *See* SOLAR RADIATION.

The region of rising temperature above 80 km (48 mi) is known as the thermosphere. The exosphere is that region of the atmosphere that is so rarefied that for many purposes collisions between molecules can be neglected; it is roughly the region above 500 km (300 mi).

The maximum temperature in the thermosphere and exosphere on a diurnal basis occurs in late afternoon and near the subsolar latitude. The diurnal maximum is about 30% greater than the minimum, which occurs in the early morning hours and at about the same latitude as the subsolar point but in the opposite hemisphere. On average, the temperature of

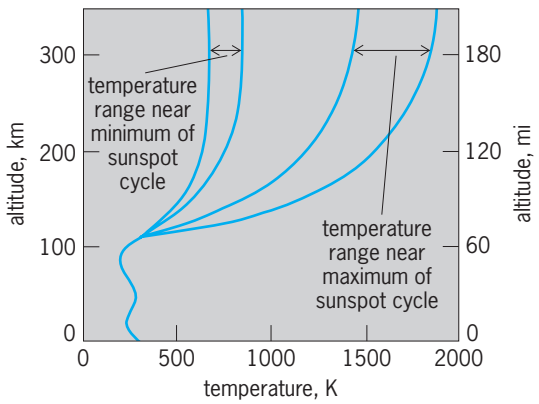


Fig. 2. Some typical temperature distributions through the atmosphere. $°F = (K \times 1.8) - 459.67$.

the exosphere over the summer pole is about 300 K (540°F) higher than over the winter pole.

Where the atmosphere is warm, it is less dense, and the pressure fall with altitude is less rapid than where the atmosphere is cold. Thus the warmest region of the upper atmosphere is also the region of highest pressure. The resultant horizontal pressure gradients tend to make the atmosphere flow away from the warm region toward the cold. While Coriolis force comes into play to modify this tendency, in the upper atmosphere there are other effects that reduce the importance of the Coriolis force. These effects are viscosity and ion drag. Viscosity acts to reduce vertical gradients in the wind field. Ion drag occurs because the ions in the ionosphere are not free to drift across magnetic field lines; when the neutral atmosphere moves, collisions between neutral atmospheric particles and ions provide a drag force on the neutral atmospheric motion. These forces cause the global-scale wind patterns in the thermosphere to be relatively simple, mainly blowing from the warmest region toward the coldest. Wind patterns have been calculated taking all these forces (including Coriolis) into account, and the calculations are in reasonable agreement with the observations, although observations are difficult to make and are rather sparse. *See* CORIOLIS ACCELERATION.

The fact that the coldest region of the thermosphere occurs at mid or low latitude rather than over the winter polar region where sunlight is entirely lacking results from downward motion and compressional heating that occurs over the winter polar region—about 10^{-1} cm/s at 50 km (30 mi), 10 cm/s at 100 km (60 mi), and 10^{2} cm/s above 150 km (90 mi). This compensates in considerable degree for the lack of solar heating there. A similar effect occurs in the vicinity of the temperature minimum; without it the minimum temperature would be much lower than it is.

An important, though sporadic, heat input in the upper atmosphere is auroral zone heating. At times of geomagnetic disturbances, important heating occurs there as a result of the inflow of energetic charged particles along the geomagnetic field lines and as a result of electric currents that cause Joule heating as they flow through the lower ionosphere. In especially strong events, the auroral zone heating completely disrupts the normal pattern of upper atmospheric circulation, and the flow in all but the lower thermosphere is from the poles to the Equator. On such occasions the auroral heating is a more important heat input to the upper atmosphere than is the absorption of solar ultraviolet radiation. More common are events where the normal circulation is only partially disrupted and both circulations are recognizable—one away from the normal diurnal temperature maximum and the other away from the poles. *See* AURORA; GEOMAGNETIC VARIATIONS; THERMOELECTRICITY.

At high latitudes, the ionosphere moves in response to electric fields imposed as a consequence of interactions between the Earth's magnetic field and the solar wind. The imposed electric field causes the ionosphere to drift in a generally antisunward direction over the polar caps (regions at higher latitudes than the auroral zone, or magnetic latitudes greater than about 68°), with a return circulation (that is, generally sunward in direction) just outside the polar caps. Although the ions are much less numerous than the neutral particles and they transfer their momentum to the neutral particles when they collide, they quickly regain their drift velocity through action of the electric field, and each ion can therefore transfer momentum many times to the neutral particles, making up for their lesser numbers by repeated collisions, each transferring momentum. Because of the persistence of ionospheric drift caused by electric fields, the neutral particles gradually pick up a drift velocity similar to that of the ions. Thus a velocity pattern similar to the polar-region ionospheric drift pattern becomes superposed on the large-scale circulation of the neutral thermosphere. Over the polar caps, the two motions are largely in the same direction, but the ionospheric circulation imposes a reverse circulation on the neutral atmosphere in a relatively narrow region just outside the polar caps. *See* ATMOSPHERIC GENERAL CIRCULATION; SOLAR WIND; WIND.

Composition. Ultraviolet radiation of suitable wavelengths can photodissociate atmospheric molecules—something of great importance in the upper atmosphere. It is even important in the stratosphere, where ozone is formed as a result of absorption by molecular oxygen of ultraviolet radiation, the important wavelengths being below 242 nanometers. Dozens of chemical reactions are actually involved in ozone chemistry, but in highly simplified terms the essential mechanism is shown by reactions (2), where M is a third body that is

$$\begin{aligned} O_2 + h\nu &\rightarrow O + O \\ O + O_2 + M &\rightarrow O_3 + M \\ O + O_3 &\rightarrow 2O_2 \end{aligned} \qquad (2)$$

required for the reaction to proceed, even though it is not changed chemically by the reaction. An equilibrium between formation and destruction of ozone is established. This basic set of reactions is important in the altitude range 15–70 km (9–42 mi); it indicates in a highly simplified way how the ozone layer is formed. *See* STRATOSPHERIC OZONE; STRATOSPHERE.

At higher altitudes where ultraviolet radiation is present that cannot penetrate as deeply as the ozone layer, the importance of oxygen photodissociation is greater. As atomic oxygen becomes more plentiful and molecular oxygen less plentiful, the main loss mechanism for atomic oxygen is reaction (3).

$$O + O + M \rightarrow O_2 + M \qquad (3)$$

At altitudes above 100 km (60 mi), recombination becomes very slow because of low concentrations and the resultant smaller probability of collisions. As a result, the loss of atomic oxygen occurs mainly by diffusion downward to the region below 100 km (60 mi) where recombination proceeds more rapidly.

Above 100 km (60 mi), the distribution of atmospheric gases becomes increasingly controlled by diffusion of the various constituents among themselves; the end result of the diffusion process in the gravitational field is the diffusive-equilibrium distribution of atmospheric constituents in which each constituent is distributed as if the others were not present. Atomic oxygen, being much lighter than molecular oxygen and nitrogen, falls off with altitude much more slowly than the molecular species. Above about 200 km (120 mi), atomic oxygen predominates over molecular oxygen and nitrogen. A similar situation with atomic nitrogen does not occur, because it photodissociates much more slowly; it requires much shorter-wave-length ultraviolet radiation (<80 nm), which is much less intense in solar radiation.

The photodissociation of oxygen above 100 km (60 mi), the diffusion downward of the resultant atomic oxygen to altitudes below 100 km (60 mi), and its recombination there into molecular oxygen provide a downward transport of chemical energy. Energy from ultraviolet radiation is converted partly into chemical form (by production of atomic oxygen) and partly into heat at the altitude where it is absorbed (the excess energy beyond that required for photodissociation being released as kinetic energy). The chemical energy is released below 100 km (60 mi) when recombination into molecular oxygen occurs.

The large-scale circulation of the thermosphere produces a rather surprising effect on atmospheric composition. In the lower part of the altitude region where atmospheric constituents are in diffusive equilibrium, horizontal winds are more effective in transporting lighter constituents than heavier ones. Because of the lesser rate of decrease in concentration with altitude for the lighter constituents, an effectively thicker layer is transported by the wind system. As a consequence, the large-scale circulation of the thermosphere transports relatively more atomic oxygen than molecular oxygen and nitrogen from the summer to the winter hemisphere and toward the winter polar region. The effect is so pronounced that it results in atomic oxygen enrichment of the winter thermosphere, in spite of the fact that the rate of photoproduction is very low in that hemisphere. This atomic oxygen enrichment of the winter thermosphere has an important effect on ionospheric chemistry, and the winter ionosphere is more highly ionized as a result of the enrichment. The enrichment effect is even stronger is the case of helium, because it is lighter than atomic oxygen. The enrichment in the case of helium is referred to as the winter helium bulge. *See* AERONOMY.

Planetary upper atmospheres. Data from other planets, most notably Venus and Mars, have considerably expanded understanding of upper atmospheric dynamics. Venus has a surface atmospheric pressure almost a hundred times greater than Earth, whereas Mars has a surface pressure about one-hundredth that of Earth. Surprisingly, the upper atmospheres of these two planets are much more similar to one another than to the upper atmosphere of Earth. This similarity is due to the absence of intrinsic magnetic fields large enough to shield the atmospheres from the solar wind (Mars may have an intrinsic field just on the borderline of providing some shielding effect), and to the very similar atmospheric compositions on the two planets (about 96% carbon dioxide in both cases).

The most detailed data that have been obtained are for Venus, mostly from the *Pioneer Venus* and *Venera 11* and *12* missions. These show that the solar wind confines the ionosphere on the sunward side of the planet, the interplanetary magnetic field providing the mechanism for the interaction. Since the solar wind is quite variable in intensity, the altitude at which the interface occurs is also variable, ranging from 200 to 1000 km (120 to 600 mi) but frequently in the vicinity of 400 km (240 mi). The daytime ionosphere tends to escape from the confinement region by flowing at high velocity (several kilometers per second) across the entire terminator from the dayside hemisphere to the night-side hemisphere, providing an important source for the maintenance of the night-time ionosphere.

The thermosphere of Venus is very cold compared to that of Earth, about 250 to 300 K (−10 to 80°F) on the day side and 100 K (−280°F) on the night side. The neutral thermospheric circulation is undoubtedly a flow from the day side to the night side, as both the horizontal neutral pressure gradient and momentum transfer from the rapidly flowing ionosphere act in this sense.

Measurements of the vertical distributions of atmospheric constituents can be interpreted in terms of the average rate of turbulent mixing in the lower thermosphere, and the result is a surprisingly large rate, much larger than on Earth. The rate is so high that there is some reason to doubt the correctness of the interpretation.

The upper atmosphere of Mars is very similar to that of Venus. The great difference in surface pressure on the two planets does not have any significance with regard to the upper atmosphere, and other parameters are rather similar. It is plausible to assume that internal gravity-wave activity on Mars might be much greater than on Venus because surface winds are greater; this is a direct result of the low density of the Martian atmosphere at the surface and its consequent low heat capacity. Winds blowing over mountains are likely sources of internal gravity waves, and the surface winds on Mars are known to be strong, at least at times. On Venus, on the other hand, surface winds are weak, as the high-density atmosphere can transfer large amounts of heat with only light winds. *See* MARS; VENUS.

Francis S. Johnson

Bibliography. A. Brekke, *Physics of the Upper Polar Atmosphere*, 1995; S. Chapman and R. S. Lindzen, *Atmospheric Tides, Thermal and Gravitational*, 1970; T. E. Cravens and A. F. Nagy, Aeronomy of the inner planets, *Rev. Geophys. Space Phys.*, 21:263–273, 1983; W. Dieminger, C. K. Hartmann, and R. Leitinger, *The Upper Atmosphere: Data Analysis*

and Interpretation, 1994; M. Ghil and S. Childress, *Topics in Geophysical Fluid Dynamics: Atmospheric Dynamics, Dynamo Theory, and Climate Dynamics*, 1987; C. Riege, *Fundamentals of Atmospheric Dynamics and Thermodynamics*, 1990; R. G. Roble, R. E. Dickinson, and E. C. Ridley, Global circulation and temperature structure of the thermosphere with high-latitude plasma convection, *J. Geophys. Res.*, 87:1599–1614, 1982.

Upsilon particles

A family of elementary particles whose first three members were discovered in 1977. The upsilon mesons, Υ, are the heaviest known vector mesons, with masses greater than 10 times that of the proton. They are bound states of a heavy quark and its antiquark. The quarks which bind to form the upsilons carry a new quantum number called beauty or bottomness, and they are called *b*-quarks or *b*. The mass of the *b*-quark is around 5 GeV. The anti-*b*-quark or $\overline{b}$ carries antibeauty, and therefore the upsilons carry no beauty and are often called hidden beauty states. Direct proof of the existence of the *b*-quark was obtained by observing the existence of *B* mesons which consist of a *b*-quark bound to a lighter quark. Twelve $b\overline{b}$-mesons have also been observed thus far. *See* QUARKS.

Heavy quarks. In addition to electric charge, all quarks carry a new nonscalar "charge," called color, which results in forces much stronger than the electric Coulomb force. There is, however, a major difference between the electric and the color forces. While the Coulomb force becomes weaker at large distances, color forces are believed to become stronger at large distances, resulting in the confinement and thus unobservability of free quarks. The theory of the color forces is known as quantum chromodynamics (QCD). *See* COLOR (QUANTUM MECHANICS); QUANTUM CHROMODYNAMICS.

A bound system of a heavy quark and its antiquark ($q\overline{q}$ or quarkonium) has very small dimensions. As a result, the color force becomes relatively weak and the quarks move with low velocity. The quark-antiquark system can therefore be described by a Schrödinger equation if the potential between the quarks is known. In a pure coulombic potential, a $q\overline{q}$ system becomes very similar to an hydrogen atom. A better analogy, however, is the case of positronium, an atomlike system consisting of an electron and a positron bound together by their mutual electric attraction. Just as for the case of positronium, which can decay into photons, the quanta of the electromagnetic field, so can the quarkonium state annihilate into gluons, the quanta of the field coupled to the color charge. According to older ideas, heavy quarkonium states should annihilate in times of the order of 10^{-24} s, while quantum chromodynamics suggests much longer lifetimes. *See* ATOMIC STRUCTURE AND SPECTRA; GLUONS; NONRELATIVISTIC QUANTUM THEORY; POSITRONIUM; QUANTUM MECHANICS.

In 1974 the first heavy vector meson was discovered. The mass of this new particle, called J/ψ, is around a 3 GeV and its width of the order of 60 keV, corresponding to a lifetime of about 10^{-19} s. The discovery of the J/ψ (a $c\overline{c}$ bound system, where *c* stands for the charm quark of mass around 1.5 GeV), was a triumph for quantum chromodynamics. *See* CHARM; J/PSI PARTICLE.

The upsilon mesons provide even better testing grounds for quantum chromodynamics. Because they contain the much heavier *b* quarks, the color force is slightly weaker than for the J/ψ case. Thus the upsilon annihilation rate was expected to be slower or, equivalently, its width to be narrower. This is indeed so: the width of the upsilon is only 40 keV while its mass is 9.4 GeV.

Potential models. Quantum chromodynamics has not yet been able to derive the exact form of the potential. It can, however, suggest its form. In particular, the strength of the potential is supposed to increase linearly at large distances and to obey a $1/r$ law at short distances, like the Coulomb potential. On this basis, the properties of the ψ and Υ families have been thoroughly calculated. The amount of experimental information is now so large that in fact the process can be reversed, and from the data the form of the potential has been extracted to great accuracy.

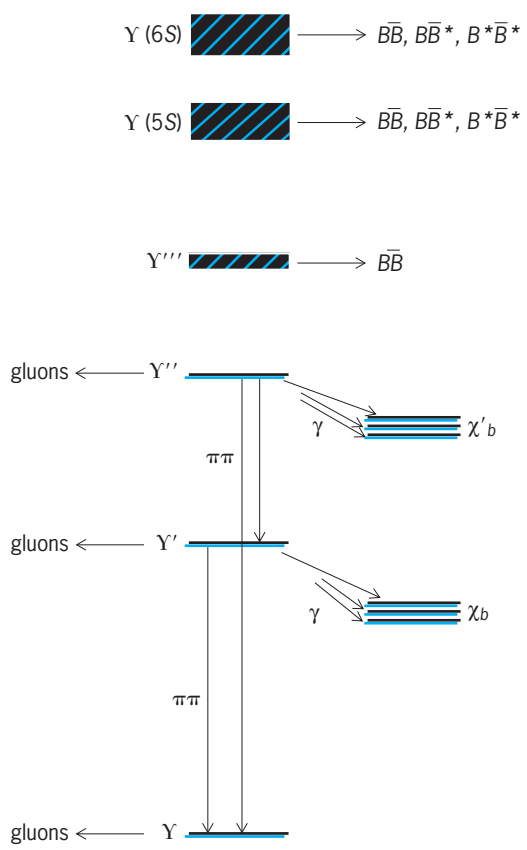
Family of states. Just as for the case of the hydrogen atom or of positronium, the $b\overline{b}$ pair can bind in many different ways, properly classified by the quantum numbers of the various states. The bound pair can have orbital angular momentum of 0, 1, 2, and so forth, the corresponding states being referred to as *S*-wave, *P*-wave, *D*-wave, and so forth. Each quark, in addition, has a spin angular momentum of $^1/_2$. The two spins can combine to give a total spin of 0 or 1. Spin-0 states, as in all quantum systems, are called singlet states, and spin-1 states, triplet states. Orbital and spin angular momentum combine to give the total angular momentum of the state. Bound states are usually classified by the principal quantum number *n*, the total angular momentum *J*, and the total spin *S* (see **table** and **illus.**).

S-wave states. While for a pure coulombic potential there exists an infinite number of bound states, the quarkonium potential requires that there should

Masses of the known members of the upsilon family

Meson name	Mass, GeV/c^2*
Υ	9.46037 ± 0.00021
Υ'	10.02330 ± 0.00031
Υ''	10.3553 ± 0.0005
Υ'''	10.5800 ± 0.0035
$\Upsilon(5S)$	10.865 ± 0.008
$\Upsilon(6S)$	11.019 ± 0.008
$\chi_{b,J=2}$	9.9132 ± 0.0006
$\chi_{b,J=1}$	9.8919 ± 0.0007
$\chi_{b,J=0}$	9.8598 ± 0.0013
$\chi'_{b,J=2}$	10.2685 ± 0.0004
$\chi'_{b,J=1}$	10.2552 ± 0.0005
$\chi'_{b,J=0}$	10.2321 ± 0.0006

*c = speed of light.



Known $b\bar{b}$ levels and their intertransitions.

be only three *S*-wave, spin-triplet, bound $b\overline{b}$ states. These are, in fact, the first three upsilon mesons discovered in 1977. The Υ mass is now known to about 20 parts per million, an accuracy close to that for the electron mass.

The fourth upsilon, Υ''', was discovered in 1980. In contrast to the first three upsilons, which have widths ranging from 40 to 20 keV, the width of the Υ''' is about 20 MeV, almost a thousand times greater, indicating that it is slightly heavier than two *B* mesons into which it decays. Direct evidence for the decay $\Upsilon''' \rightarrow B + \overline{B}$ was obtained by the CESR experiments, by observing the beta decay of the *B* mesons.

The fifth and sixth radial excitations of the triplet *S*-wave $b\overline{b}$ system were observed in 1984. These mesons are usually called $\Upsilon(5S)$ and $\Upsilon(6S)$.

P-wave states. Six spin-triplet, *P*-wave bound $b\bar{b}$ states, called χ_b, are expected to exist. These states cannot be produced in e^+e^- annihilations because they have positive parity and charge conjugation. They can, however, be produced in a two-step process given by expressions (1) and (2). In this way

$$e^+e^- \rightarrow \Upsilon \tag{1}$$

$$\Upsilon \rightarrow \chi_b + \gamma \tag{2}$$

the lowest spin-triplet *P*-wave state can be reached, after emission of a photon, from the Υ', and the second *P*-wave state can be reached via decays of the Υ''. Because the triplet *P*-wave states have total spin of 1 and also orbital angular momentum 1, the total angular momentum of these states can be, according to the rules of quantum mechanics, 0 or 1 or 2. The spin-orbit interaction shifts the mass of states with different total angular momentum, resulting in six different states: $\chi_{b,J=2}$, $\chi_{b,1}$, $\chi_{b,0}$, $\chi'_{b,2}$, $\chi'_{b,1}$, and $\chi'_{b,0}$. The three χ'_b were discovered in 1982 by observing the decays $\Upsilon'' \rightarrow \chi + \gamma$, $\chi \rightarrow \Upsilon' + \gamma$, and $\chi \rightarrow \Upsilon + \gamma$, and the three χ_b were discovered in the same manner in 1983. Later the χ_b states were also observed by other experiments.

In addition to electromagnetic transitions between Υ's and χ's, transitions between Υ's, with emission of two pions, have also been observed. *See* ELEMENTARY PARTICLE; MESON.

Paolo Franzini

Bibliography. P. Franzini and J. Lee-Franzini, Upsilon resonances, *Annu. Rev. Nucl. Part. Phys.*, 33:1–29, 1983; G. Kane, *Modern Elementary Particle Physics*: *The Fundamental Particles and Forces*, 1993; L. M. Lederman, The upsilon particle, *Sci. Amer.*, 239(4):72–80, October 1978; Review of particle properties, *Phys. Rev. D*, 50:1173–1825, 1994.

Upwelling

The phenomenon or process involving the ascending motion of water in the ocean. Vertical motions are an integral part of ocean circulation, but they are a thousand to a million times smaller than the horizontal currents. Vertical motions are inhibited by the density stratification of the ocean because with increasing depth, as the temperature decreases, the density increases, and energy must be expended to displace water vertically. The ocean is also stratified in other properties; for example, nutrient concentration generally increases with depth. Thus even weak vertical flow may cause a significant effect by advecting nutrients to a new level. *See* GEOPHYSICAL FLUID DYNAMICS.

There are two important upwelling processes. One is the slow upwelling of cold abyssal water, occurring over large areas of the world ocean, to compensate for the formation and sinking of this deep water in limited polar regions. The other is the upwelling of subsurface water into the euphotic (sunlit) zone to compensate for a horizontal divergence of the flow in the surface layer, usually caused by winds.

The upwelling of abyssal water is very slow (meters per year) and is inferred from numerical models, the distribution of properties, and the requirement that the circulation in the deep ocean balance the sinking of cooled surface water in polar regions. The residence time of the abyssal water, that is, the time from formation and sinking in polar regions to upwelling after flowing at depth through the world ocean, is hundreds of years. It had been thought that upwelling of abyssal water resulted from the downward mixing of heat across the thermocline over vast areas of the low- and mid-latitude oceans, buoying up the old abyssal waters so that they could be replaced

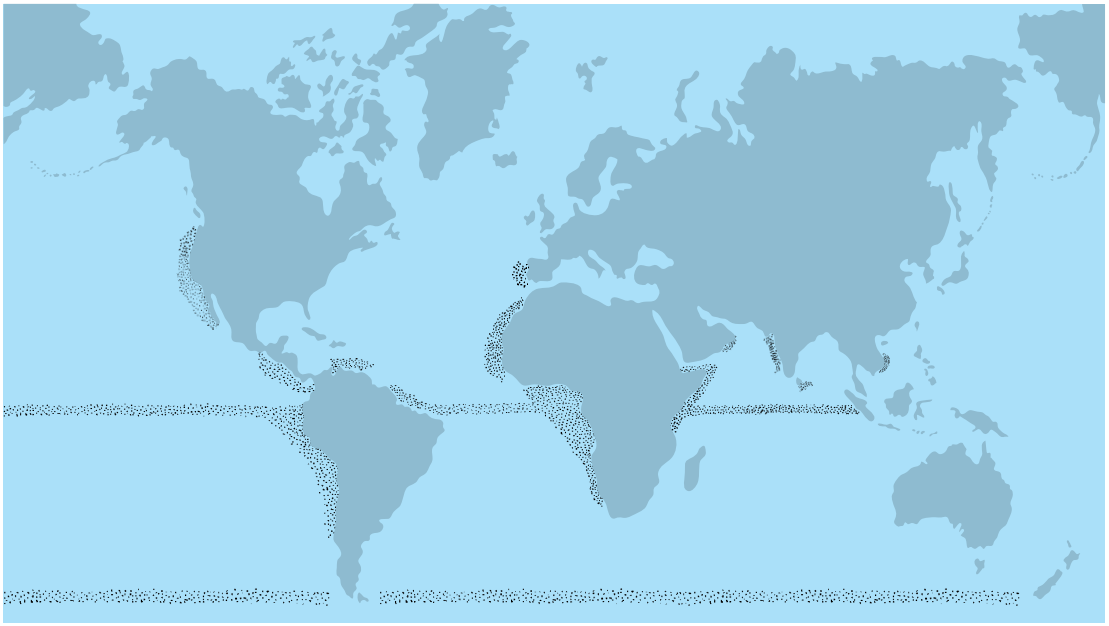


Fig. 1. Regions of major upwelling are stippled.

by new deep water flowing from the polar regions. It is now apparent that mixing in the deep ocean is insufficient to account for the amount of abyssal upwelling, and that surface forcing by the winds may be responsible.

Wind-induced divergence in the surface currents drives the more rapid (meters per day) upwelling of water from a hundred meters' depth into the euphotic zone. This upwelling can have an immediate effect on the ecology of the ocean and the world's economy. If the mechanism causing the divergence continues, cold and nutrient-rich subsurface water will continue to upwell to the surface layer and spread out horizontally. Eventually, sinking or downwelling will occur to balance this upwelling, but this may occur far from the area of upwelling. Regions of persistent upwelling are characterized by relatively cool surface water that is rich in nutrients compared with typical oceanic surface water. The upwelling flux of nutrients increases the region's fertility. Unusually high biological productivity is associated with these upwelling regions, which are often obvious in satellite images of sea surface temperature or chlorophyll distributions. *See* OCEAN CIRCULATION.

Occurrence. Upwelling may occur anywhere in the ocean where a mechanism exists to cause a divergence in the surface layer flow. It occurs consistently in both the equatorial zone and the Antarctic circumpolar ocean but is probably most conspicuous along the western coasts of the continents (**Fig. 1**). At low and middle latitudes in both hemispheres, the surface waters within a hundred kilometers of the eastern boundaries of the ocean (the west coasts) are cooler (by several degrees Celsius) and more productive biologically (by a factor of 10 or more) than water farther seaward at the same latitudes as a result of upwelling near the coast.

Wind-induced upwelling was explained qualitatively by V. W. Ekman (1905), who studied the influence of the wind on the surface currents. He showed that the effects of the Earth's rotation (the Coriolis force) and frictional forces (due to turbulence) resulted in the net movement of the surface layer being directed 90° to the right of the wind in the Northern Hemisphere and 90° to the left in the Southern Hemisphere. This wind-induced transport, called Ekman transport, is confined to a layer only several tens of meters deep. Quantitatively, the Ekman transport can be calculated from the wind stress (the frictional drag force exerted by the wind on the sea surface) divided by the Coriolis parameter (twice the Earth's rotation rate multiplied by the sine of the latitude). *See* CORIOLIS ACCELERATION; WIND; WIND STRESS.

Because of semipermanent midocean atmosphere high-pressure systems, the prevailing winds are generally equatorward along the western coasts of the continents. Thus the Ekman transport is directed away from the coasts, and this one-sided divergence necessitates an upwelling of subsurface water into the surface layer near the coast (**Fig. 2**). The upwelled water comes from depths of 50 to a few hundred meters and reaches the surface layer in a relatively narrow band extending only a few tens of kilometers from the coast. The intensity of coastal upwelling must balance the Ekman transport directed offshore since the latter must be replaced by upwelling near the coast. The upwelling index used by the U.S. National Marine Fisheries Service is exactly this. Typical Ekman transport values off Oregon and California during summer are about 1000 kg/s per meter of coastline. If this water upwells through a band extending about 10 km from the coast, the vertical velocity would be only about 0.0001 m/s (about 10 m per day). This compares with typical horizontal currents of 0.1–1 m/s (about 10–100 km per day).

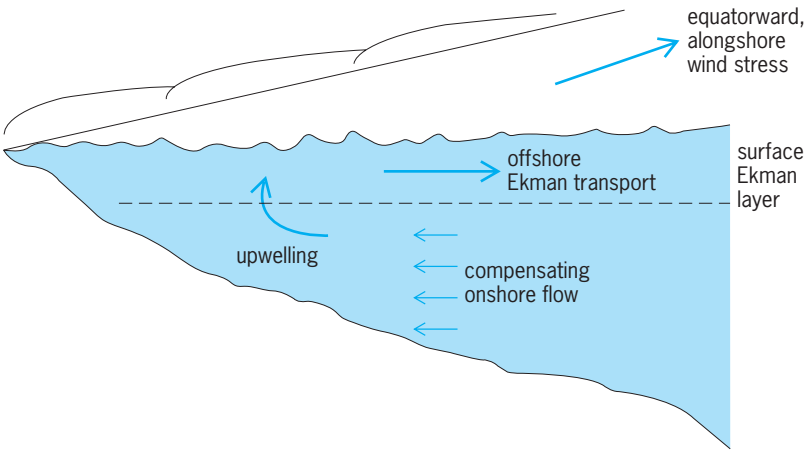


Fig. 2. Schematic diagram of coastal upwelling.

Upwelling at the Equator also results from the wind-induced divergence of the surface layer, analogous to that in coastal upwelling. The trade winds in the equatorial regions blow westward, and thus the surface water is transported away from the Equator on either side of it because the Ekman transport is to the right of the wind in the Northern Hemisphere and to the left in the Southern Hemisphere. The divergence of the Ekman transport near the Equator results in the upwelling of cool nutrient-rich water.

There is also strong upwelling in the Arabian Sea (northwestern Indian Ocean) off the east coasts of Somalia, Saudi Arabia, and Oman during May to September when the monsoon winds blow from the southwest, nearly parallel to the coast, resulting in offshore Ekman transport. The monsoon wind increases to maximum offshore, causing a further divergence in the surface flow, augmenting the coastal upwelling. The sea surface temperature at the Arabian coast may drop below 20°C, while temperatures 1000 km seaward are nearly 30°C. During the winter the winds blow in the opposite direction and no upwelling occurs. *See* MONSOON METEOROLOGY.

Very near the Antarctic continent the winds blow westward, but farther to the north the winds blow eastward. Between these winds, there is a region of diverging flow in the surface layer, which causes nutrient-rich water to upwell. The eastward winds blow unimpeded around the Antarctic continent and, with the Ekman transport directed to the left of the wind in the Southern Hemisphere, the Ekman transport diverges and causes upwelling. Large blooms of plankton occur, providing the food for krill that are eaten by whales, which is why the ocean around Antarctica was once a major whaling ground. The water, although dense, is weakly stratified, and thus water may be upwelled from great depth using relatively little energy. This wind-induced upwelling may help drive the abyssal ocean's overturning circulation. *See* ANTARCTIC OCEAN.

Effects. As a result of coastal upwelling, the water near the coast is denser, and this sets up an alongshore (equatorward) geostrophic current off the west coasts. The boundary between upwelled water and the offshore oceanic surface water often has a region of sharp temperature and density contrast (a front). The alongshore geostrophic current in this region may be very strong (a coastal jet) because of the sharp density gradient. The trajectory of an upwelling parcel of water is relatively flat; the parcel moves horizontally with the currents along the coast as it gradually rises toward the surface. The upwelled water spreads offshore and equatorward, and the area affected by coastal upwelling may eventually extend 100 km or more offshore. The flow pattern is not simple, and the boundary between upwelled and oceanic water may meander, with filaments of cold water being apparent in satellite images of the sea surface temperature along the west coasts of the continents. Along the west coasts, low stratus clouds or fog is common due to the cooling of air near the coast as it comes into contact with the cool upwelled water. The cool ocean temperatures along the California coast contrast considerably with the warmer (by as much as 10°C) temperatures along the eastern United States seaboard at the same latitudes. The early explorers to the west coasts of Africa and South America were surprised by the anomalously cool surface waters near the coast, and the Conquistadores reportedly made good use of the upwelling by hanging flagons of wine in the sea to cool while in Peruvian ports.

Biologically, coastal upwelling regions are economically important. Their primary productivity is high because nutrients are supplied to the euphotic zone by the upwelling process. This provides a base for the short efficient food chains that characterize coastal upwelling regions. It is estimated that half of the fish caught commercially are from coastal upwelling regions, which constitute only 0.1% of the ocean's area. The coast of Peru is legendary for the large population of guano birds and the anchoveta fishery on which the birds feed. Because of the relative narrowness of the coastal upwelling regions, high fish concentrations can be efficiently harvested, making the species vulnerable to overfishing. The short food chain and low species diversity make the coastal upwelling ecosystem especially vulnerable to natural environmental variations. Evidence for climate variations can be found in the marine sediments, containing the undissolved remains of high productivity. *See* BIOLOGICAL PRODUCTIVITY; MARINE ECOLOGY; MARINE FISHERIES; MARINE SEDIMENTS; SEAWATER FERTILITY.

Variability. Upwelling varies with changing winds, with latitude, and, like the weather, with the changing seasons. In some regions, there is a seasonal reversal of the winds as atmospheric pressure systems shift. An example is the west coast of North American north of about 40° latitude (northern California), where the winds from late fall until early spring are predominantly from the southwest, resulting in onshore Ekman transport and the suppression of upwelling. Off southern California, Mexico, and Peru, the upwelling continues throughout the year, but the intensity varies with latitude and season. However,

off Peru, coastal upwelling is affected by an occasional catastrophic variation—El Niño.

El Niño originally referred to the occasional appearance of excessively warm water along the coast of Peru and was once thought to be due to weakened coastal upwelling. It is now known that El Niño is a larger-scale phenomenon originating in the western equatorial Pacific that, among other effects, causes the warm and nutrient-depleted surface layer to deepen greatly off Ecuador and Peru. Local coastal upwelling continues under undiminished coastal winds during El Niño, but cool nutrient-rich water is out of reach of the upwelling process and only warm nutrient-depleted water flows into the euphotic zone, collapsing its usually high productivity. At the same time, the winds in the equatorial Pacific decrease and equatorial upwelling ceases or diminishes greatly, and the equatorial surface waters warm. The resulting change in ocean-atmosphere interaction in the equatorial region affects the weather worldwide. *See* EL NIÑO; MARITIME METEOROLOGY; SEAWATER; TROPICAL METEOROLOGY.

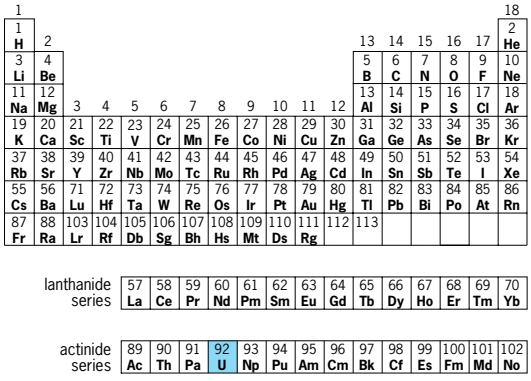
Robert L. Smith

Bibliography. A. E. Gargett, Physical process and the maintenance of nutrient-rich euphotic zones, *Limnol. Oceanog.*, 36:1527–1545, 1991; S. J. Lentz, The surface boundary layer in coastal upwelling regions, *J. Phys. Oceanog.*, 22:1517–1539, 1992; W. J. Schmitz, Jr., On the interbasin-scale thermohaline circulation, *Rev. Geophys.*, 33:151–173, 1995; C. P. Summerhayes et al. (eds.), *Upwelling in the Ocean: Modern Processes and Ancient Records*, Wiley, 1995; J. Robert Toggweiler, The ocean's overturning circulation, *Phys. Today*, pp. 45–50, November 1994.

Uranium

A chemical element, symbol U, atomic number 92, atomic weight 238.03. The melting point is 1132°C (2070°F) and the boiling point is 3818°C (6904°F). Uranium is one of the actinide series. *See* ACTINIDE ELEMENTS; PERIODIC TABLE.

Uranium in nature is a mixture of three isotopes: ^{234}U, ^{235}U, and ^{238}U. Uranium is believed to be concentrated largely in the Earth's crust, where the average concentration is 4 parts\break per million (ppm).



The total uranium content of the Earth's crust to a depth of 15 mi (25 km) is calculated to be 2.2×10^{17} lb (10^{17} kg); the oceans may contain 2.2×10^{13} lb (10^{13} kg) of uranium. Several hundred uranium-containing minerals have been identified, but only a few are of commercial interest. *See* RADIOACTIVE MINERALS; URANINITE.

Because of the great importance of the fissile isotope ^{235}U, rather sophisticated industrial methods for its separation from the natural isotope mixture have been devised. The gaseous diffusion process, which in the United States is operated in three large plants (at Oak Ridge, Tennessee; Paducah, Kentucky; and Portsmouth, Ohio) has been the established industrial process. Other processes applied to the separation of uranium include the centrifuge process, in which gaseous uranium hexafluoride is separated in centrifuge cascades, the liquid thermal diffusion process, the separation nozzle, and laser excitation. *See* ISOTOPE (STABLE) SEPARATION.

Uranium is a very dense, strongly electropositive, reactive metal; it is ductile and malleable, but a poor conductor of electricity. Many uranium alloys are of great interest in nuclear technology because the pure metal is chemically active and anisotropic and has poor mechanical properties. However, cylindrical rods of pure uranium coated with silicon and canned in aluminum tubes (slugs) are used in production reactors. Uranium alloys can also be useful in diluting enriched uranium for reactors and in providing liquid fuels. Uranium depleted of the fissile isotope ^{235}U has been used in shielded containers for storage and transport of radioactive materials. *See* NUCLEAR FUELS; NUCLEAR REACTOR.

Uranium reacts with nearly all nonmetallic elements and their binary compounds. Uranium dissolves in hydrochloric acid and nitric acid, but nonoxidizing acids, such as sulfuric, phosphoric, or hydrofluoric acid, react very slowly. Uranium metal is inert to alkalies, but addition of peroxide causes formation of water-soluble peruranates. *See* URANIUM METALLURGY.

Uranium reacts reversibly with hydrogen to form UH_3 at 250°C (482°F). Correspondingly, the hydrogen isotopes form uranium deuteride, UD_3, and uranium tritide, UT_3. The uranium-oxygen system is extremely complicated. Uranium monoxide, UO, is a gaseous species which is not stable below 1800°C (3270°F). In the range UO_2 to UO_3, a large number of phases exist. The uranium halides constitute an important group of compounds. Uranium tetrafluoride is an intermediate in the preparation of the metal and the hexafluoride. Uranium hexafluoride, which is the most volatile uranium compound, is used in the isotope separation of ^{235}U and ^{238}U. The halides react with oxygen at elevated temperatures to form uranyl compounds and ultimately U_3O_8.

Fritz Weigel

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; J. J. Katz et al. (eds.), *The Chemistry of the Actinide and Transactinide Elements*, 3d ed., 2006; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., CRC Press, 2004.

Uranium metallurgy

The processing treatments for the production of uranium concentrates and the recovery of pure uranium compounds, as well as the conversion chemistry to produce uranium metal and the processes employed for preparing uranium alloys.

Ore processing. The procedures to recover uranium from its ores are numerous, because of the great variety in the nature of uranium minerals and associated materials and the wide range of concentration in the naturally occurring ores. Recovery of uranium requires chemical processing; however, preliminary treatment of the ore may involve a roasting operation, a physical or chemical concentration step, or a combination of these.

Roasting can bring about significant chemical change on its own, but moreover it can improve the filtering and settling characteristics of an ore for subsequent processing. Only a limited number of deposits are sufficiently rich in uranium-bearing minerals and have appropriate physical characteristics to make a preliminary physical concentration of the uranium a feasible operation. Chemical concentration is most common as a preliminary treatment.

In general, one of two leaching treatments—acid leaching and carbonate leaching—is used as the initial step in chemical concentration. The choice depends on the nature of the ore, which largely determines the efficiency and the cost of the process employed.

Acid leaching is carried out universally with sulfuric acid. The finely ground ore is treated with the dilute acid, and usually an oxidant is added to make sure that all the uranium goes into solution as the uranyl ion. When an ore is carbonate-leached, the action of sodium carbonate on the uranium forms water-soluble $[UO_2(CO_3)_3]^{4-}$ ion. This solution is stabilized by adding bicarbonate and by keeping the uranium in the oxidized state. From either the acid or carbonate leaching process, the uranium can be recovered as ammonium uranate or sodium uranate (yellow cake) by precipitation with ammonia or with sodium hydroxide. *See* LEACHING.

Recovery of pure uranium. The concentrate, whether obtained by chemical or physical means, is treated chemically to give a uranyl nitrate solution that can be further purified by solvent extraction. The impurities remain in the aqueous phase, while the uranium is extracted into the organic phase. Formerly, ether or hexone was used as an extractant; presently, tributyl phosphate (TBP) is favored. From the organic phase, the uranium is stripped by means of water, and the high-purity uranium can be recovered as nitrate crystals or by precipitation from the solution. The nitrate serves as the starting material for other compounds, such as the oxides. In large-scale processing, the nitrate is decomposed thermally to give uranium trioxide, UO_3, which is subsequently reduced with hydrogen to form uranium dioxide, UO_2. Uranium tetrafluoride, UF_4 (green salt), is prepared by treating UO_2 with hydrogen fluoride (HF) gas. *See* SOLVENT EXTRACTION.

Metal reduction. Uranium metal can be obtained from its halides by fused-salt electrolysis or by reduction with more reactive metals. The reaction of UO_2 with calcium yields metal of fair quality.

The electrolysis of UF_4 in a fused-salt bath of sodium chloride and calcium chloride has been used to prepare tons of high-purity metal. The metal is deposited as small granules on a molybdenum cathode. In the calcium reduction of UO_2, the ingredients of the charge are mixed and placed in an inert atmosphere and are heated to about 1830°F (1000°C) to form finely divided uranium and calcium oxide. Calcium hydride can be employed instead of calcium in this process. The metal from either electrolysis or the calcium-UO_2 process is leached, pressed, and sintered or melted to give solid uranium. *See* ELECTROLYSIS.

The largest tonnages of good-quality uranium have been produced by metallothermic reduction of finely divided UF_4 with calcium or magnesium in steel bombs lined with fused dolomitic oxide (Ames process). The charge, consisting of an intimate mixture of UF_4 with the reductant metal in granular form together with a suitable booster (usually calcium plus iodine), is placed into the reduction bomb, the lid is bolted down, and the bomb is heated to ignition temperature. In the case of calcium, the reaction heat is such that the charge can be ignited at room temperature, and the products are sufficiently molten to allow separation of metal and slag in the reduction vessel. For magnesium, the whole bomb has to be preheated in a heat-soaking pit. The heat generated in the reduction reaction is sufficient to melt both metal and slag, and the molten metal collects by gravity as a pool under the molten slag, where it

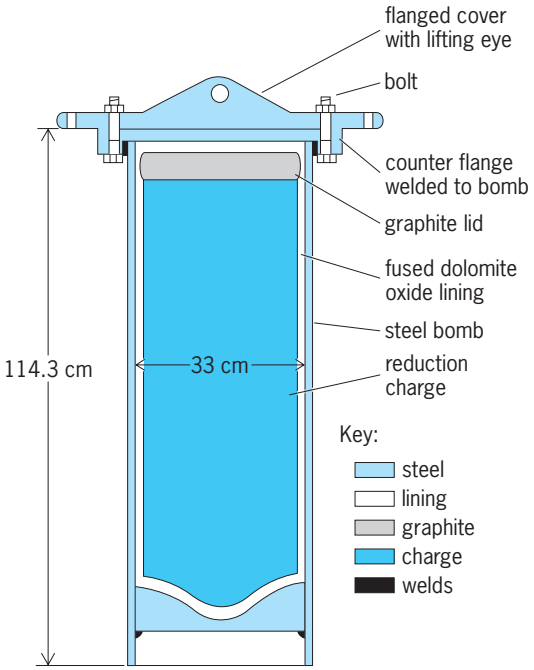


Fig. 1. Standard Derby bomb for reduction of UF_4 with magnesium by the Ames process. Capacity is 317.9 lb (144.2 kg) of uranium. 1 cm = 0.4 in.

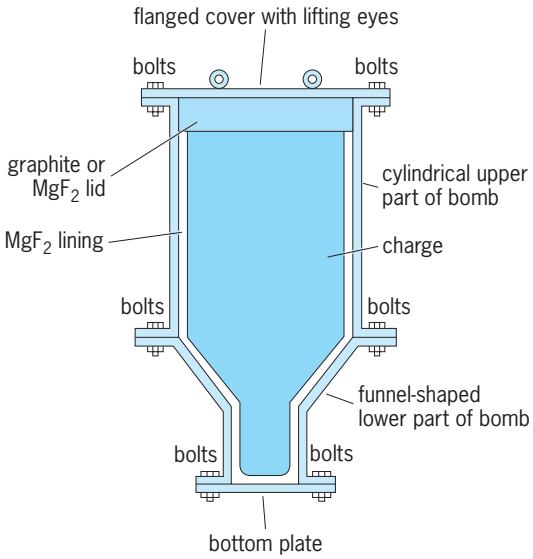


Fig. 2. Dingot bomb for large charges.

remains during solidification. The shape of the metal ingot (also referred to as a biscuit) depends on the shape of the reduction bomb. A flat, pancake-shaped ingot, which has to be remelted for reshaping, is obtained in the standard bomb (**Fig. 1**), while reduction in a Dingot bomb (**Fig. 2**) yields a so-called direct-reduction ingot (dingot), which is shaped in such a way that it may undergo further metallurgical treatment without remelting. The Dingot process is particularly suited for large-scale reductions (110 lb or 50 kg to 1.7 tons or 1.5 metric tons). The metal obtained in the metallothermic reduction process is quite pure. Metal obtained by reduction in a standard bomb has to be recast by vacuum melting in a graphite crucible and casting into graphite molds. The metal can be fabricated by conventional means with due consideration for its chemical reactivity and allotropic transformations. The low-ductility modification of uranium is usually avoided in fabrication processes. In the metallurgy of ^{233}U and ^{235}U, strict weight and configuration limitations must be imposed to avoid criticality.

Alloys. Uranium alloys are prepared by fusing the components together. All procedures following conventional metallurgical techniques, however, may have to be carried out inside inert-gas glove boxes because many alloys are attacked by oxygen or moisture. *See* NUCLEAR FUELS; NUCLEAR FUELS REPROCESSING; URANIUM. F. Weigel

Bibliography. R. C. Merritt, *Extractive Metallurgy of Uranium*, 1971, reprint 1979; Organization for Economic Cooperation and Development, *Uranium Resources: Production and Demand*, 1988.

Uranus

The first planet to be discovered with the telescope and the seventh in the order of distance from the Sun. It was found accidentally by W. Herschel in England on March 13, 1781. Herschel's home-made telescope was good enough to show that this object was not starlike; Uranus appeared as a fuzzy patch of light, not a point. At first he thought it was a comet, but subsequent calculations of the orbit demonstrated that Uranus was actually a planet, about twice as far from the Sun as Saturn and therefore in an orbit that agreed almost exactly with the prediction of the Titius-Bode relation for planetary distances. *See* PLANET.

The planet and its orbit. The main orbital elements are the semimajor axis (mean distance to the Sun) of 1.78×10^9 mi (2.87×10^9 km); the eccentricity of 0.047; the sidereal period of 83.75 years; the orbital velocity of 4.25 mi/s (6.8 km/s); and the inclination of orbital plane to ecliptic of 0.8°.

These characteristics indicate a normal, well-behaved orbital motion. Therefore it is surprising to discover that the obliquity (inclination of rotational axis to orbit plane) of Uranus is 98°, exceeded only by the obliquities of Pluto and Venus. This means that the axis is almost in the plane of the orbit, and thus the seasons on Uranus are very unusual. During summer in one hemisphere, the pole points almost directly toward the Sun while the other hemisphere is in total darkness. Forty-two years later, the situation is reversed. Thus, "day" and "night" for an observer at the north or south pole of Uranus each last more than 40 years. Recent years of solstice (pole toward Sun) are 1944 and 1985, and recent and future years of equinox (equator toward Sun) are 1966 and 2007. The south pole was almost directly facing the Sun when the *Voyager 2* spacecraft sped past the planet in January 1986. The actual period of rotation of Uranus is 17.24 h. *See* PLUTO; SPACE PROBE; VENUS.

The maximum apparent equatorial diameter of the disk of this distant planet is about 4.″1 (the Moon has an apparent diameter of 31′). The corresponding linear equatorial diameter is 31,763 mi (51,118 km). The mass is 14.5 times the mass of the Earth, a value that can be well determined from the observed motions of the satellites. The corresponding mean density is 1.27 g/cm^3, which is greater than that of Saturn even though Uranus is smaller than Saturn. This means that Uranus is richer than Saturn (or Jupiter) in elements heavier than hydrogen and helium but not nearly so rich in these elements as Earth (mean density 5.5). The same is true for Neptune; thus these two planets constitute a distinct subgroup in the outer solar system when compared with Jupiter and Saturn. *See* JUPITER; NEPTUNE; SATURN.

The apparent visual magnitude of Uranus at mean opposition, that is, when closest to Earth, is +5.5. Therefore, this planet is just visible to the unaided eye in a dark sky when its position among the stars is known. The corresponding value of the albedo, or reflectivity, is 0.5, the same as Saturn and Jupiter, and slightly higher than Neptune (0.4). The temperature that Uranus would assume in simple equilibrium with incident solar radiation at this distance is about 55 K (−360°F). This is essentially identical with the value found by direct measurement. There is thus no evidence for the existence of an internal energy

source as observed for Jupiter, Saturn, and Neptune. The temperature measured by *Voyager 2* was the same at both the sunlit and dark poles. Thermal radiation has been detected from the lower atmosphere of Uranus, indicating that the temperature increases with depth but much more slowly than in the case of the other major planets. At a wavelength of 21 cm, a temperature of 250 ± 60 K (9 ± 90°F) has been measured. *See* ALBEDO.

Through the telescope, Uranus appears as a small, slightly elliptical blue-green disk. This appearance is confirmed in the nearly featureless pictures of the planet obtained by *Voyager 2*. Uranus owes its characteristic aquamarine color to the relatively high proportion of methane in its atmosphere. This gas absorbs the orange and red wavelengths from incident sunlight, scattering back blue-green light to an observer. The methane abundance is 20 to 30 times the amount corresponding to a solar distribution of the elements, in agreement with the proportion of heavy elements in the planet deduced from its average density. In striking contrast, the proportion of helium to hydrogen, the two most abundant gases in the planet's atmosphere, is essentially equal to that observed in the Sun.

Like Jupiter and Saturn, Uranus exhibits zonal winds that are parallel to the equator, despite the unusual orientation of the planet's rotational axis. These winds appear to increase in speed with increasing latitude, unlike the winds on Jupiter and Saturn, which are strongest at the equator. They move in the same direction as the rotation of Uranus, producing a circulation pattern like that on Neptune. The causes of this circulation and the way it maintains the two poles at the same temperature have not yet been determined.

Uranus has a magnetic field equivalent to that of a bar magnet offset from the planet's rotational axis by one-third of the planetary radius and tilted by 60°. (On Earth the field is inclined only 11° to the rotational axis and offset by 0.07 of the radius.) This means that the poles of the magnetic field on Uranus occur at subtropical latitudes instead of near 80°, as on Earth. This field orientation is similar to that of Neptune and must reflect a similar (but not yet understood) mechanism for field generation on these two planets. The strength of the magnetic field on Uranus is five times that of Earth, but varies considerably over the planet. This field traps belts of electrons and protons (like Earth's Van Allen belts) in a magnetosphere whose unusual configuration is determined by the orientations of the planet's magnetic and rotational axes. Electrons and protons from these Van Allen belts bombard the upper atmosphere of Uranus and give rise to an aurora, as they do on Earth, but this aurora accurs in the subtropics instead of at high latitudes. *See* AURORA; MAGNETOSPHERE; PLANETARY PHYSICS; VAN ALLEN RADIATION.

Satellites. Before the *Voyager 2* encounter, only five satellites of Uranus had been discovered. They form a remarkably regular system with low orbital eccentricities and inclinations close to the plane of the planet's equator. The *Voyager* cameras found 11 additional small satellites closer to the planet, including two that may serve as gravitational "shepherds" for the outermost ring (see **table**). Two additional inner satellites were found from Earth in 2003. Like Jupiter, Saturn, and Neptune, Uranus has more distant, irregular satellites (moons that move in orbits with high inclinations or eccentricities). Two of these unusual moons were discovered with the Palomar 200-in. (5-m) telescope in 1997, three more were discovered

Satellites of Uranus*

Satellite	Mean distance from Uranus, 10^3 mi	Mean distance from Uranus, 10^3 km	Sidereal period, days	Radius, mi	Radius, km	Albedo (reflectivity)	Density, g/cm^3‡
Cordelia†	30.9	49.8	0.34	8	13		
Ophelia†	33.4	53.8	0.38	9	15		
Bianca	36.8	59.2	0.44	14	22		
Cressida	38.4	61.8	0.46	20	32		
Desdemona	38.9	62.6	0.47	19	30		
Juliet	40.0	64.4	0.49	26	42		
Portia	41.1	66.1	0.51	34	55		
Rosalind	43.5	70.0	0.56	19	30		
Belinda	46.8	75.3	0.62	21	34		
Puck	53.4	86.0	0.76	48	78	0.07	
Miranda	80.8	130	1.41	147	236	0.3	1.2
Ariel	119	191	2.52	360	580	0.4	1.7
Umbriel	165	266	4.14	364	585	0.2	1.4
Titania	271	436	8.71	490	789	0.3	1.7
Oberon	362	583	13.5	473	761	0.2	1.6
Caliban§	4474	7200	579	19	30	0.07¶	
Stephano§	4971	8000	677	6	10	0.07¶	
Trinculo§	5284	8504	759	3	5	0.07¶	
Sycorax§	7581	12,200	1283	37	60	0.07¶	
Prospero§	10,066	16,200	1962	9	15	0.07¶	
Setebos§	10,229	17,500	2209	9	15	0.07¶	

*Only satellites with well-determined orbits as of March 1, 2004, are listed.
†Shepherding satellites for ring.
‡Eccentricity of 0 is a circle, 1.00 is a parabola.
§Outer, retrograde satellites. Radii are derived by assuming albedos of 0.07. Densities are unknown.
¶Assumed.

in 1999, and four more since then. Only the first six to be found have reliable orbits, but the orbits of the others are being improved and the search for additional satellites continues.

The spacecraft was able to obtain pictures of the surfaces of all five of the larger satellites and resolved features on Puck as well. For reasons that are not yet understood, there is increasing evidence of internal geological activity on the larger satellites with decreasing distance from the planet. Miranda, the smallest but also the closest, revealed the most spectacular diversity of surface features (**Fig. 1**). Impact craters are present on all of these objects as well.

With densities of approximately 1.5 g/cm^3, these satellites are presumably composed of a mixture of abundant ices (dominated by water ice) and some rocky material. The generally low reflectivities and the deposits of dark material observed in some locations on the satellites' surfaces are ascribed to the presence of carbon-rich compounds. These compounds are more neutral in reflectivity than the dark material found on Saturn's satellite Iapetus, however, suggesting that more than one type of organic material is present in the outer solar system.

Rings. A system of ten rings can be observed around Uranus. They are designated by a mixture of Greek letters and Arabic numerals (**Fig. 2**). Since the rotational axis of Uranus is almost in the plane of the planet's orbit around the Sun, the rings are nearly perpendicular to this plane, which made it possible to discover nine of them when they passed in front of a star.

The *Voyager* cameras added a tenth ring, lambda, between delta and epsilon. All of these rings are remarkably narrow. The widest is epsilon, whose mean width of only 36 mi (58 km) is still six to seven times larger than that of the next widest rings. Like the small satellites, the rings are as dark as coal, with reflectivities of only 3–5%. (The Earth's Moon has a reflectivity of 11%.) The spaces between these narrow rings are often thinly populated with dust. Measurements of the diminution of starlight by the rings using a photometer on the spacecraft as well as measurements of the effect of the rings on radio transmission from the spacecraft itself suggest an absence of small particles within the epsilon ring. The origin and means of confinement of these rings are topics of ongoing research.

Cosmogony. It appears that all four giant planets have approximately equal masses of heavy elements in their interiors, equivalent to 10–20 Earth masses. The great difference between Jupiter and Saturn on the one hand, and Uranus and Neptune on the other, can then be understood in terms of differences in their atmospheres. The outer two planets acquired less gas from the original solar nebula when they formed, perhaps because the nebula was simply thinner at this large distance from its center. Since the nebular gases were predominantly hydrogen and helium, Uranus and Neptune began with a smaller endowment of these two light elements and thus ended up with mean densities larger than Jupiter and Saturn. The excess methane remarked on above was produced by the core, which was unable to trap the two lightest gases. The hydrogen and helium are therefore present in a solarlike ratio to one another.

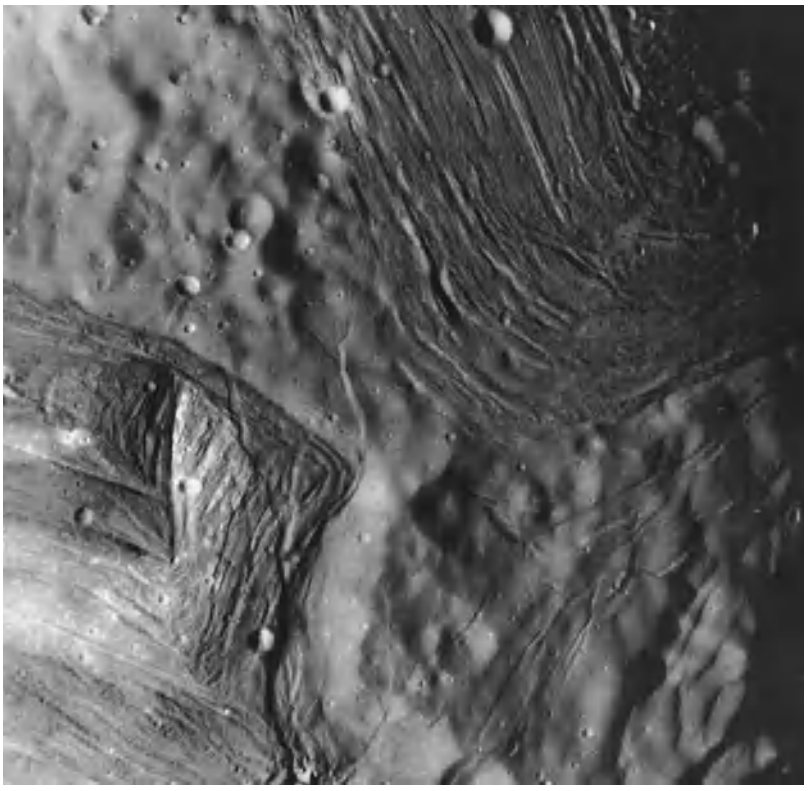


Fig. 1. Miranda viewed from *Voyager 2*. Features as small as 2000 ft (600 m) across are visible. The two regions of grooved terrain are remarkable, showing evidence of large-scale geological activity on this tiny satellite.

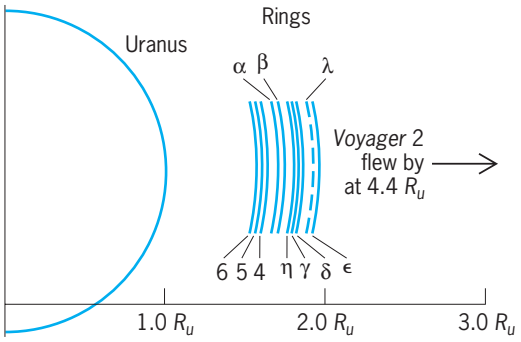


Fig. 2. Diagram of Uranus ring system, showing one hemisphere of Uranus at the left and segments of the 10 narrow rings. The rings all lie between 1.5 and 2 times the radius of Uranus (R_u) from the planet's center.

The large obliquity of Uranus and its lack of an internal heat source remain unsolved problems. The obliquity is commonly attributed to the impact by a large body during the late phases of core formation, but a detailed theory has not yet been developed.

Tobias C. Owen

Bibliography. J. K. Beatty, C. C. Petersen, and A. Chaikin (eds.), *The New Solar System*, 4th ed., Cambridge University Press, 1999; R. Greenberg and A. Brahic (eds.), *Planetary Rings*, University of Arizona Press, 1984; G. Hunt, *Uranus and the Outer Planets*, Cambridge University Press, 1982; D. Morrison

and T. Owen, *The Planetary System*, 3d ed., Addison-Wesley, 2003; Reports of *Voyager 2* encounter with Uranus, *Science*, 233:39–109, 1986; D. A. Rothery, *Satellites of the Outer Planets: Worlds in Their Own Right*, 2d ed., Oxford University Press, 1999; F. W. Taylor, *The Cambridge Photographic Guide to the Planets*, Cambridge University Press, 2001.

Urban climatology

The branch of climatology concerned with urban areas. These locales produce significant changes in the surface of the Earth and the quality of the air. In turn, surface climate in the vicinity of urban sites is altered. The era of urbanization on a worldwide scale has been accompanied by unintentional, measurable changes in city climate.

Controlling factors. The process of urbanization changes the physical surroundings and induces alterations in the energy, moisture, and motion regime near the surface. Most of these alterations may be traced to causal factors such as air pollution; anthropogenic heat; surface waterproofing; thermal properties of the surface materials; and morphology of the surface and its specific three-dimensional geometry—building spacing, height, orientation, vegetative layering, and the overall dimensions and geography of these elements. Other factors that must be considered are relief, nearness to water bodies, size of the city, population density, and land-use distributions (**Table 1**).

TABLE 1. Climatic alterations produced by cities

Element	Compared to rural environs
Contaminants	
Condensation nuclei	10 times more
Particulates	10 times more
Gaseous admixtures	5–25 times more
Radiation	
Total on horizontal surface	0–20% less
Ultraviolet: winter	30% less
summer	5% less
Sunshine duration	5–15% less
Cloudiness	
Clouds	5–10% more
Fog: winter	100% more
summer	30% more
Precipitation	
Amounts	5–15% more
Days with <5 mm (0.2 in.)	10% more
Snowfall: inner city	5–10% less
lee of city	10% more
Thunderstorms	10–15% more
Temperature	
Annual mean	0.5–3.0°C (0.9–5.4°F) more
Winter minima (average)	1–2°C (1.8–3.6°F) more
Summer maxima	1–3°C (1.8–5.4°F) more
Heating degree days	10% less
Relative humidity	
Annual mean	6% less
Winter	2% less
Summer	8% less
Wind speed	
Annual mean	20–30% less
Extreme gusts	10–20% less
Calm	5–20% more

*After H. E. Landsberg, *The Urban Climate*, Academic Press, 1981.

Urban heat island. In general, cities are warmer than their surroundings, a fact documented over a century ago. They are islands or spots on the broader, more rural surrounding land. Thus, cities produce a heat island effect on the spatial distribution of temperatures (**Fig. 1**). The timing of a maximum heat island is followed by a lag shortly after sundown (Fig. 1*c–e*), as urban surfaces, which absorbed and stored daytime heat, retain heat and affect the overlying air. Meantime, rural areas cool at a rapid rate (Fig. 1*d*).

There are a number of energy processes that are altered to create warming, and various features lead to those alterations (**Table 2**). City size, the morphology of the city, land-use configuration, and the geographic setting (such as relief, elevation, and regional climate) dictate the intensity of the heat island, its geographic extent, its orientation, and its persistence through time. Individual causes for heat island formation are related to city geometry, air pollution, surface materials, and anthropogenic heat emission. There are two atmospheric layers in an urban environment. Besides the planetary boundary layer (PBL) outside and extending well above the city, the first of these layers is the urban boundary layer. This layer is due to the spatially integrated heat and moisture exchanges between the city and its overlying air (**Fig. 2**). The surface of the city corresponds to the level of the urban canopy layer. Fluxes across this plane comprise those from individual units of the second layer, the urban canopy layer, such as roofs, canyon tops, trees, lawns, and roads, integrated over larger land-use divisions (for example, suburbs).

Energy and moisture balance. An understanding of changes of climate in the city is achieved by investigating the energy and moisture regimes in the urban and rural locales.

The urban energy balance is the sum of the net all-wave radiation (inputs and outputs of solar and longwave radiation) and the energy released through anthropogenic activities. This sum is calculated by adding the latent heat flux (evapotranspiration), sensible (turbulent) heat flux, storage heat flux (into surface, buildings, and so forth), and the net heat advection (horizontal heat transfer). The latent heat flux is the product of the latent heat of vaporization and the mass equivalent of the latent heat flux. These values apply to the top of a volume that extends to sufficient depth that vertical heat and water exchange are negligible. The fluxes are determined at the top of the volume (for example, in watts per square meter).

The urban water balance is the sum of the precipitation, the piped water supply, and the water released through anthropogenic activities. This sum is calculated by adding the parameters of runoff, the measured change in storage for the period of interest, the moisture advection, and the mass equivalent of the evapotranspiration.

The moisture exchanges of cities differ in several ways from those occurring in rural settings. Precipitation can increase in, or downwind of, a

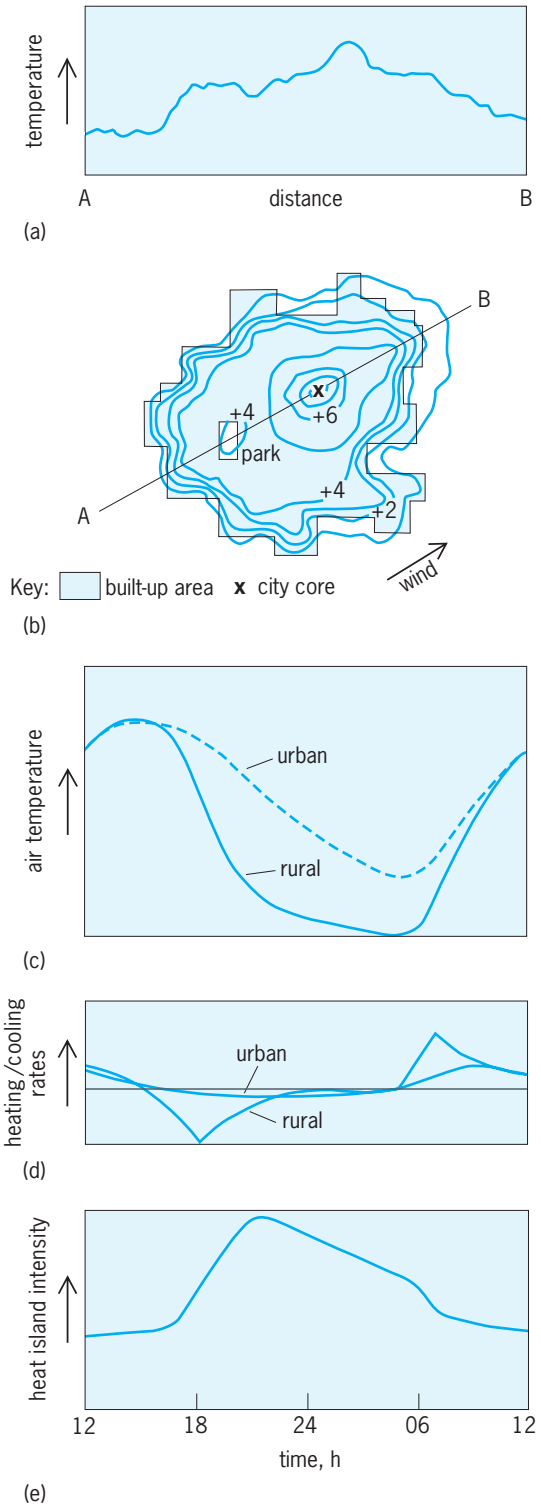


Fig. 1. Example of heat island magnitude, spatial arrangement, and timing during the day under "ideal" (calm, clear) weather. (*a*) The cross-sectional changes in temperature across the heat island spatial pattern in relation to (*b*) the plan outline of a city, (*c*) air temperature in an urban and rural locale during 24 h, (*d*) air heating and cooling rates per hour, and (*e*) resulting heat island intensity. Vertical scale units are around 2°C for *a*, *c*, and *d* and 2°C/h for *e*. Values at the contours in *b* represent temperatures in °C associated with built-up area in comparison to rural areas. °F = (°C × 1.8) + 32. (*After T. R. Oke, The energetic basis of the urban heat island, Quart. J. Roy. Meteorol. Soc., 108:1–24, 1982*)

large urban area (Table 1), and significant increases occur in the warm season, which is associated with more local convective processes. The city environment is aerodynamically rougher, contains more pollution condensation nuclei, promotes more thermal and mechanical turbulence, has an effect on the formation of convective clouds, and enhances precipitation. Development of a clear characterization of the role of urban places in affecting precipitation is complicated by the topographic settings of cities.

Since the piped water supply and the water released through anthropogenic activities are extra sources of moisture in a city, there is a net gain of moisture by the process of urbanization. Water loss through mass exchange of water (evapotranspiration) is considered to be less in a city because of a "waterproofing effect." This effect is brought about by replacement of moisture surfaces with more impervious ones with the result that the ratios of latent heat flux to net all-wave radiation have been found to exceed expectations. For example, in a temperate city, evapotranspiration has been shown to constitute 38% of the losses of the annual external water balance and 81% of the losses of the summer water balance. Assuming moisture advection to be negligible, runoff would increase in cities. Many studies have indicated increased runoff in urban environments as well as altered hydrographic characteristics of urban streams. Changes in land use have been shown to have four effects on the hydrology of an urban area: (1) peak flow is increased; (2) total runoff is increased; (3) water quality is lowered; and (4) hydrologic amenities such as the appearance of river channel and esthetic impressions are lowered. Discharge regimes are a function of the percentage of impervious surface area in cities and percentage of area served by storm sewers. *See* HYDROGRAPHY; HYDROLOGY.

Since for a given period evapotranspiration and change in water storage are typically lower, dewpoint is usually lower in the daytime in urban areas. At night, a humidity island may result in the city because of extra transfer of evapotranspiration, and water released through anthropogenic activities. The largest unknown in analyses of water balance for cities is a basic understanding of transfers of evapotranspiration from complex suburban and urban landscapes. The measured spatial variability of evapotranspiration is dependent on the complex source areas of moisture transfer to fixed observed points across the urban and suburban environment.

Air movement in cities. Urban areas affect the speeds and directions of winds on both a micro- and mesoscale. The aerodynamic roughness of cities is larger than that of surrounding rural environments, thus inducing more frictional drag. The heat island effect influences the pressure field and vertical stability of the air over the city. The city influence on wind is also a function of the gradient air motion above the city. Many studies have shown that as general winds in the region increase, the intensity of the heat island diminishes. However, the built-up environment of the city induces mechanical turbulence and may

TABLE 2. Suggested causes of the urban heat island

Energy processes that are altered to create warming	Features of urbanization underlying energy balance changes
Urban canopy layer	
Increased absorption of short-wave radiation	Canyon geometry—increased surface area and multiple reflection
Increased long-wave radiation from the sky	Air pollution—greater absorption and reemission
Decreased long-wave radiation loss	Canyon geometry—reduction of sky view factor
Anthropogenic heat source	Building and traffic heat losses
Increased sensible heat storage	Construction materials—increased thermal admittance
Decreased evapotranspiration	Construction materials—increased "waterproofing"
Decreased total turbulent heat transport	Canyon geometry—reduction of wind speed
Urban boundary layer	
Increased absorption of short-wave radiation	Air pollution—increased aerosol absorption
Anthropogenic heat source	Chimney and stack heat losses
Increased sensible heat input-entrainment from below	Canopy heat island—increased heat flux from canopy layer and roofs
Increased sensible heat input-entrainment from above	Heat island, roughness—increased turbulent entrainment

*After T. R. Oke, The energetic basis of the urban heat island, *Quart. J. Roy. Meteorol. Soc.*, 108:1–24, 1982.

speed up winds across the area by some 30%. For conditions of light regional airflow, development of an intense heat island and its effects on vertical stability and pressure field variations will tend to promote quite variable patterns of air motion in the urban area. In rural areas, air is more stable, and little motion occurs. In cities, the thermal influences induce increases in turbulence and wind speed.

Horizontal advection of air, and the heat content of that air, further complicates a complete understanding of heat fluxes from sectors of urban environments. This advective component of motion and heat transfer may play a role in the enhanced dispersion potential of pollution in urban regions. The large spatial variability of heat fluxes across urban areas has been postulated to affect plume spread. *See* ATMOSPHERIC GENERAL CIRCULATION; WIND.

Heat island influences affect the mixing height of the atmosphere—a measure of vertical pollutant mixing above the urban canopy layer. Basically, a domelike lifting of inversions occurs over the city center, with subsidence of the inversion on the perimeter of the urban region. Aircraft overflights and remote detectors measure atmospheric mixing and inversion conditions with devices such as lidar and Doppler radar. *See* DOPPLER RADAR; LIDAR.

Satellite climatology. Satellite platforms are used to study the distribution of heat islands and the finer instantaneous spatial details of the heat and moisture at the surface. It is possible that uncritical acceptance of remotely sensed data can lead to erroneous conclusions. In the search for in-depth physical understanding of the causes of heat islands, many questions arise. For example, what is the nature of the true surface "seen" by space platform sensors? What is the relation between temperatures interpreted by the satellite and the three-dimensional surface temperatures across the urban environment? Satellite estimates of radiant surface temperatures provide a "snapshot" of the entire city region at a resolution not made possible previously—a large advantage in assessing the spatial dimensions of a heat island. However, the resultant heat island interpretation is specific to the satellite observation system, such as the instantaneous time of the satellite pass over an area (not often at the time of maximum heat island formation) and the view angle of the resultant temperature pattern of surfaces seen only by the satellite radiometer. *See* CLIMATE MODIFICATION; CLIMATOLOGY; SATELLITE METEOROLOGY.

Anthony J. Brazel

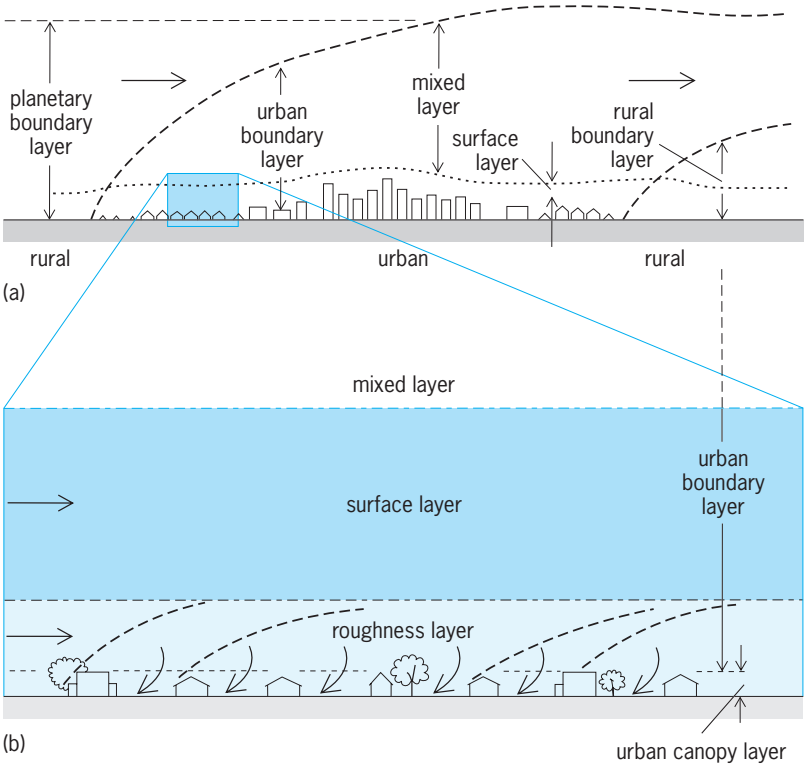


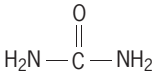
Fig. 2. Diagrams showing exchange of heat and moisture between a city and its overlying air. (*a*) Idealized arrangement of boundary-layer structures over a city; horizontal arrows indicate general direction of prevailing wind. (*b*) A blowup of a section of the rural layer in *a*, showing the mixed, surface, and roughness layers, including eddies of air shown by curved arrows, and individual mixed layers associated with land conditions, indicated by broken curved lines. (*After T. R. Oke, The urban energy balance, Prog. Phys. Geog., 12:471–508, 1988*)

Bibliography. J. K. S. Ching, J. F. Clarke, and J. M. Godowitch, Modulation of heat flux by different scales of advection in an urban environment, *Boundary-Layer Meteorol.*, 25:171–191, 1983; C. S. B. Grimmond and T. R. Oke, An evapotranspiration-interception model for urban areas, *Water Resources Res.*, 27:1739–1755, 1991; H. E. Landsberg, *The Urban Climate*, 1981; T. R. Oke,

The energetic basis of the urban heat island, *Quart. J. Roy. Meteorol. Soc.*, 108:1–24, 1982; T. R. Oke, The urban energy balance, *Prog. Phys. Geog.*, 12:471–508, 1988; M. Roth and T. R. Oke, Satellite-derived urban heat islands from three coastal cities and the utilization of such data in urban climatology, *Int. J. Remote Sens.*, 10:1699–1720, 1989.

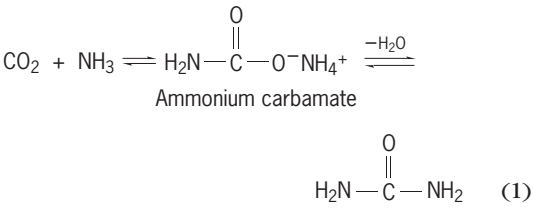
Urea

A colorless crystalline compound, formula CH_4N_2O, melting point 132.7°C (270.9°F). Urea is also known as carbamide and carbonyl diamide, and has numerous trade names as well. It is highly soluble in water and is odorless in its purest state, although most samples of even high purity have an ammonia odor. The diamide of carbonic acid, urea has the structure below.



Urea occurs in nature as the major nitrogen-containing end product of protein metabolism by mammals, which excrete urea in the urine. The adult human body discharges almost 50 g (1.8 oz) of urea daily. Urea was first isolated in 1773 by G. F. Rouelle. By preparing urea from potassium cyanate (KCNO) and ammonium sulfate (NH_4SO_4) in 1828, F. Wöhler achieved a milestone, the first synthesis of an organic molecule from inorganic starting materials, and thus heralded the modern science of organic chemistry. *See* NITROGEN; PROTEIN METABOLISM.

Preparation. Prior to World War II, urea was prepared commercially by the hydrolysis (reaction with water) of calcium cyanamid ($CaCN_2$). The development during the 1940s of large-scale high-pressure reaction vessels made of nickel-chromium alloy allowed the Kolbe process to become the principal means for producing urea. This process, first developed in the 1870s, involves the reaction of ammonia (NH_3) and carbon dioxide (CO_2) at 100 atm (10 megapascals) and 150–200°C (302–392°F) in the presence of a catalyst such as an acid or a base. This reaction forms ammonium carbamate (NH_4COONH_4), which upon further heating undergoes dehydration (loss of water) to yield urea. The overall process is shown in reaction (1). This is a very important industrial chemical process, but there are numerous other methods for producing urea.

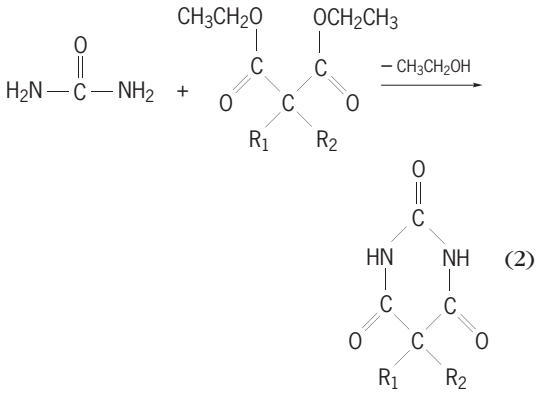


Uses. Because of its high nitrogen content (46.65% by weight), urea is a popular fertilizer. About three-fourths of the urea produced commercially is used for this purpose. After application to soil, usually as a solution in water, urea gradually undergoes hydrolysis to ammonia (or ammonium ion) and carbonate (or carbon dioxide). Another major use of urea is as an ingredient for the production of urea-formaldehyde resins, extremely effective adhesives used for laminating plywood and in manufacturing particle board, and the basis for such plastics as malamine. *See* FERTILIZER; UREA-FORMALDEHYDE RESINS.

Other uses of urea include its utilization in medicine as a diuretic. In the past, it was used to reduce intracranial and intraocular pressure, and as a topical antiseptic. It is still used for these purposes, to some extent, in veterinary medicine and animal husbandry, where it also finds application as a protein feed supplement for cattle and sheep. Urea has been used to brown baked goods such as pretzels. It is a stabilizer for nitrocellulose explosives because of its ability to neutralize the nitric acid that is formed from, and accelerates, the decomposition of the nitrocellulose. Urea was once used for flameproofing fabrics. Mixed with barium hydroxide, urea is applied to limestone monuments to slow erosion by acid rain and acidic pollutants.

Chemistry. Because of its ability to bond to itself in an orderly fashion by hydrogen bonding, urea reversibly forms inclusion complexes with straight-chained hydrocarbons having seven or more carbon atoms. These complexes consist of the hydrocarbon encased in a cylindrical wrapping of urea molecules. They are isolable as waxy solids, and can be used for separating straight-chained from branched hydrocarbons in a mixture. *See* CLATHRATE COMPOUNDS; HYDROGEN BOND.

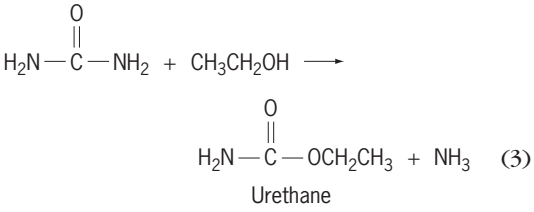
Urea has numerous applications in synthetic organic chemistry because its nitrogen and oxygen atoms are nucleophilic and its carbon atom is electrophilic. One important example of a nucleophilic reaction of urea is its reaction with acid chlorides or esters to form *N*-acyl derivatives known as ureids. Examples of ureids are acetylurea ($CH_3CO—NHCONH_2$) and the diureid diacetylurea ($CH_3CO—NHCONH—COCH_3$). The six-membered cyclic diureids formed from urea and various substituted malonic acid esters [reaction (2)] are known as barbiturates. The barbiturates are potent, addictive



$R_1 = R_2 = H$: barbituric acid
$R_1 = R_2 = CH_3CH_2$: veronal
$R_1 = CH_3CH_2, R_2 = C_6H_5$: phenobarbital

hypnotics (sleep inducers) used in medicine. *See* BARBITURATES; ESTER.

An example of an electrophilic reaction of urea is its reaction with ethyl alcohol, under pressure, to form urethane (ethyl urethane, ethyl carbamate), as shown in reaction (3). Urethane is also synthesized



by heating urea nitrate, the salt formed from the reaction between urea and nitric acid, with ethyl alcohol. Urethane was once used as a mild hypnotic, as a topical bactericide, and in the treatment of leukemia. Its toxicity precludes its continued use for these purposes. Urethane is currently used primarily as an organic solvent and as a reagent for organic synthesis. *See* ORGANIC CHEMISTRY; ORGANIC SYNTHESIS.

Robert D. Walkup

Bibliography. L. F. Fieser and M. Fieser, *Reagents for Organic Synthesis*, vol. 1, 1967; F. A. Lowenheim and M. K. Moran, *Faith, Keyes and Clark's Industrial Chemicals*, 4th ed., 1975; B. Meyer, *Urea-Formaldehyde Resins*, 1979; T. W. Solomons, *Organic Chemistry*, 7th ed., 1999.

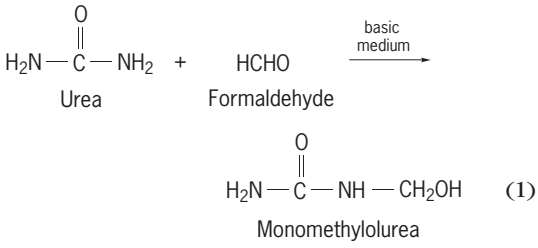
Urea-formaldehyde resins

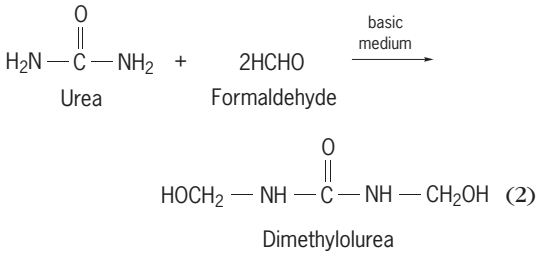
The condensation products obtained by the reaction of urea or melamine with formaldehyde. Resinous condensation products of formaldehyde with other nitrogen-containing compounds, for example, aniline and amides, also belong to this group of resins but have gained only limited utility. Resins derived from the condensation of formaldehyde with urea, melamine, aniline, and *p*-toluene sulfonamide are discussed below.

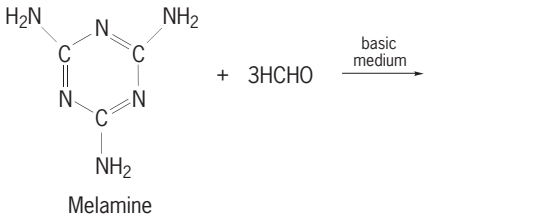
Urea- and melamine-formaldehyde resins. Because the amino resins possess an excellent combination of physical properties and can be easily fabricated in a variety of colors, they are used as adhesives, laminating resins, molding compounds, paper and textile finishes, and surface coating. Methods of utilizing the resins are generally similar to methods employed for several other condensation resins, such as the phenolic or epoxy resins. First, intermediate condensation resins are prepared; then, after compounding, the combinations are cured to yield a finished product.

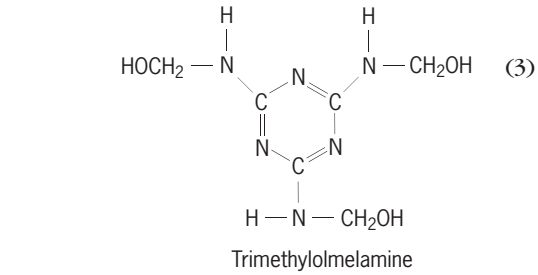
The intermediate condensation products are prepared by the reaction of urea or melamine formaldehyde under neutral or mildly alkaline conditions to form mono-, di-, or polymethylol derivatives, with reactions (1)–(3).

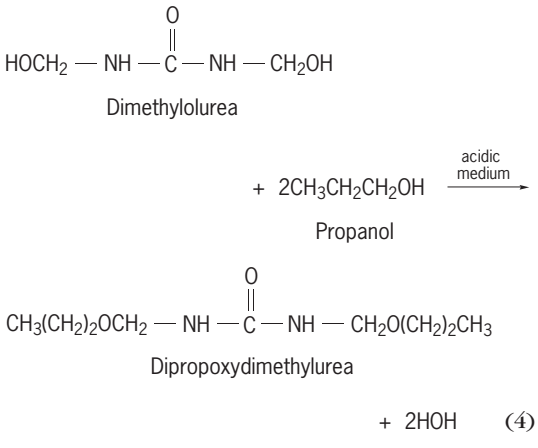
If a soluble resin is desired for surface coatings, the dimethylol derivatives are treated with an alcohol in an acidic medium to form ethers, reaction (4).





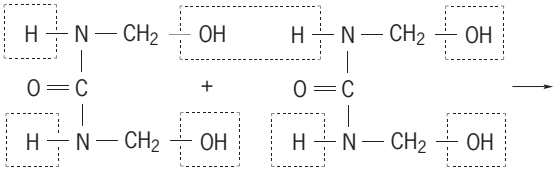


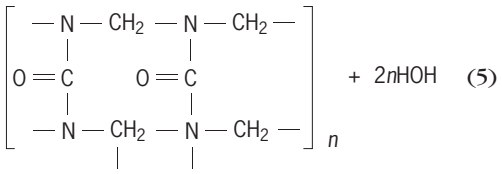




For other purposes, the low-molecular-weight methylol derivatives can then be compounded with fillers, catalysts, plasticizers, and pigments, and cured, usually by the application of heat and pressure. During the curing operation, the hydroxyl portion of a methylol group condenses with a hydrogen atom to yield water. Such a condensation can be

represented by reaction (5) for the cross-linking of methylol urea.





(5)

Cross-linked polymer

Other reactions are also involved in curing. For example, the methylol groups can react with hydroxyl groups of alcohols (added as plasticizers), clays, or cellulose (added as fillers). Methylol groups can also react with each other to form ether linkages.

For immediate curing, a free acid can be used as a cross-linking catalyst. However, for preparation of a stable molding powder that can be stored, a salt or ester that will liberate its corresponding acid at the high temperature used for molding usually serves as the catalyst.

As may be seen from reactions (1)–(5), the completion of the curing reaction depends on the elimination of as much water as possible. Although most of the water is lost as vapor, a hygroscopic filler, such as cellulose, asbestos, or certain clays, is added to absorb the last traces of moisture. In addition, the filler can also react with some of the methylol groups of the resin intermediates.

Although the fabrication and applications of both the urea- and melamine-type resins are generally similar, there are several important differences in properties. The melamine-type resins are more resistant to water, marring, and heat than the urea-type resins, but are somewhat more expensive.

Curing in a mold results in translucent or opaque products, depending on the nature of the fillers and pigments used. The ease with which color can be introduced into amino resins has made it possible to make very attractive articles, such as dinnerware, buttons, appliance cases, handles, and knobs. In applications where resistance to heat and scratching is important, for example, in dinnerware, the melamine resins are preferred. *See* PLASTICS PROCESSING.

Another important application of both resins is in adhesives, especially for lamination of furniture and plywood. Depending on the resin and catalyst used, adhesives that can be cured under a variety of conditions may be formulated. Melamine resins are especially valuable as adhesives for laminates of paper or fabrics, including glass cloth. *See* ADHESIVE.

Foams were developed for use in packaging and as thermal and acoustical insulation.

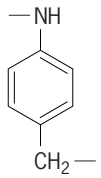
Soluble urea-formaldehyde and melamine-formaldehyde resins may be used to impregnate cloth or to treat paper. Curing of the resins in the cloth results in improved qualities, such as resistance to creasing or, in the case of melamine resins, in resistance to shrinkage in woods. Curing of the resins in paper results in an improvement in the strength of the paper when wet.

Although unmodified resins are common in the applications cited, modified resins or blends with other resins are often useful in some cases. In combination with alkyd resins, the ether derivatives, especially the ethers of the melamine resins, are components in the formulation of baking enamels that are hard and resistant to water and detergents. Resins modified by the use of an alkyl-substituted urea or melamine are more flexible than the unmodified resin and are thus useful in coating compositions. Blends of urea- or melamine-type resins with resins derived from resorcinol and formaldehyde are sometimes employed in adhesive compositions and as binders for the sawdust or wood chips used in the manufacture of particulate boards.

Another application of blends is the use of urea-type resin blends as impregnants for fabric used in the manufacture of minimum-care or wash-and-wear cotton goods. *See* TEXTILE CHEMISTRY.

Aniline- and sulfonamide-formaldehyde resins. In addition to urea and melamine, other compounds containing $—NH_2$ groups can condense with formaldehyde to form methylol derivatives which are capable of further reaction. Although the corresponding resins have received limited attention, two examples are discussed here, aniline and *p*-toluenesulfonamide.

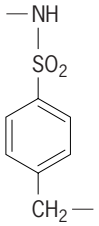
In a neutral medium, the reaction of aniline with formaldehyde yields a methylol derivative that exists as a cyclic trimer. When the trimer is heated, further condensation results in a resin with the probable monomer unit



Some cross-linking through the phenyl rings may also occur.

Because of their resistance to the absorption of water, the resins have been used in electrical applications, such as insulation or panels, where their natural brown color is not objectionable.

Similarly, resins can be prepared by curing methylol derivatives resulting from the condensation of *p*-toluenesulfonamide with formaldehyde. The probable monomer unit is



The resulting polymers are less colored than the aniline-type resins and have been employed in surface coatings. *See* FORMALDEHYDE; POLYMERIZATION; UREA. John A. Manson

Bibliography. J. R. Fried, *Polymer Science and Technology*, 2d ed., 2003; G. Odian, *Principles of Polymerization*, 4th ed., 2004; M. P. Stevens, *Polymer Chemistry: An Introduction*, 3d ed., 1998.

Uredinomycetes (rust)

An order of fungi known as plant rusts that belong to the division (phylum) Basidiomycota. In nature, all 7000 species are obligate parasites of many vascular plant species. They cause diseases known as rust on numerous cultivated crops such as asparagus, onions, beans, snapdragons, and sunflowers. Ferns, fern allies, pines and their relatives, and many other trees are also attacked. Each rust species infects one or just a few closely related plant host species.

The body of a rust fungus consists of numerous microscopic, threadlike, branching hyphae that grow inside the tissues and between the cells of the host plant. Specialized feeding structures, haustoria, grow from these hyphae into the host cells of the plant and provide nourishment for the hyphae. Hyphal cells are binucleate. Depending upon the rust species and environmental conditions, up to six different kinds of spore-producing structures, the sori, may be produced by one rust species. These sori, which are usually no more than 0.5 mm in width, may be powdery, waxy, or crustlike, and whitish, yellow, orange, brown, or blackish. An infection may produce many sori, and each sorus may produce thousands of spores. The spores are mostly wind-borne. The urediniospores of *Puccinia graminis* (the rust species that causes severe epidemics on wheat, some other cereals, and other grasses) may be disseminated for several thousand kilometers.

Rust teliospores that are produced in sori called telia are essential elements for classification. More than 125 different forms of telia have been discovered. These variations in the telia are the basis for the classification of the genera of rusts. Teliospores consist of one or more specialized probasidial cells. During the development of each of these cells, nuclei fuse and meiosis occurs. This meiosis results in a septate metabasidium from which four haploid meiospores (basidiospores) are formed. Basidiospores are forcibly ejected from the tips of their tiny stalks and are wind-disseminated. After landing on a susceptible host, under proper conditions basidiospores produce new infections. In cold regions, a rust may survive the winter as dormant, thick-walled, dark-colored teliospores. In regions of long, hot, dry summers, similar teliospores may survive the summer. In many rust species, teliospores mature and produce basidiospores with no or little dormancy, especially in the tropics. Basidiospore infections often give rise to sexual reproductive structures: the spermogonia that produce spermatia, the aecial initials (female structures), and multiple zygotes (aeciospores). Asexual spores (urediniospores) are formed by most rust species and are the stage of rusts most often found.

Although most rust species require only one host species to complete their life cycles (autoecious rusts), some of the best-known rusts require two taxonomically unrelated hosts (heteroecious rusts). For example, the rust *Cronartium ribicola* produces spermogonia and aecia on large galls on stems of pine trees, and uredinia and telia on leaves of currants and gooseberries. *Puccinia graminis* produces spermogonia and aecia on leaves of barberries, and uredinia and telia on wheat and many other grasses.

Rust fungi occur on all continents except Antarctica. They infect host plants in more than 250 vascular plant families. Rust species are especially abundant in the grass, legume, potato, coffee, and sunflower families. Under natural conditions, most rust species are in balance with their hosts and do little harm. Under agricultural ecosystems, many rusts are feared pathogens of important crops, including cereals, many beans and peas, peaches and plums, pine and eucalyptus trees, and coffee. Rust diseases are controlled most effectively by breeding resistant host varieties. This often requires the search for resistance in closely related wild relatives of crop plants followed by cross breeding to introduce wild-type resistance genes into domesticated varieties. Fungicides are used for some rusts. In the case of some heteroecious rusts, eradication of one of the hosts (the noneconomic one) has been successful. Some rust species are used to aid in biological control of weeds. *See* BASIDIOMYCOTA; EUMYCOTA; FUNGI; PLANT PATHOLOGY. Joe F. Hennen

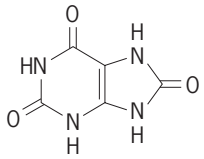
Bibliography. G. B. Cummins and Y. Hiratsuka, *Illustrated Genera of Rust Fungi*, rev. ed., 1983; L. J. Littlefield, *Biology of the Plant Rusts: An Introduction*, 1981; K. J. Scott and A. K. Chakravorty (eds.), *The Rust Fungi*, 1982.

Uric acid

An end product of purine metabolism in humans and higher primates. Uric acid is excreted as such in the urine, and it is very poorly soluble in aqueous solutions, causing it to crystallize when concentrations of the compound in the urine are abnormally high. This leads to crystalluria (excretion of crystals in urine), hematuria (blood in the urine), infection, or urinary tract stones. Other mammals do not experience these problems, as they express an enzyme, uricase, which catalyzes the conversion of uric acid to allantoin, which is highly soluble. A nonprimate exception is the Dalmatian dog, in which high rates

of excretion of uric acid are a consequence of defective renal tubular reabsorption; thus, the uric acid is excreted before it can be oxidized. In birds and reptiles, uric acid is an excretory end product of the metabolism of proteins.

Biosynthesis and metabolism. Purines are nitrogenous bases comprising a nine-membered ring formed by the fusion of a six-membered pyridine ring and a five-membered imidazole ring (see below). Exam-



Uric acid

ples of purine bases include adenine and guanine, the building blocks of ribonucleic acid (RNA) and deoxyribonucleic acid (DNA); hypoxanthine; and xanthine, the immediate catabolic precursor of uric acid. As they are found in DNA, purine nucleotides are composed of the purine base, a ribose molecule, and one to three phosphate moieties. Purine nucleotides such as inosinic acid and hypoxanthine ribose phosphate (inosine 5′-monophosphate, or IMP) are synthesized de novo from small molecules, such as glutamine, bicarbonate, formate, glycine, and aspartic acid. They may also be recovered from the free bases by salvage pathways (that is, the reclaiming of purines resulting from the breakdown of cellular DNA and RNA), catalyzed by enzymes such as hypoxanthine phosphoribosyltransferase (HPRT) and adenine phosphoribosyltransferase. The enzyme which catalyzes the conversion of xanthine to uric acid, xanthine oxidase, also catalyzes the formation of xanthine from hypoxanthine. *See* NUCLEIC ACID; NUCLEOTIDE; PURINE.

Hyperuricemia. Uric acid may be found in large amounts in blood (hyperuricemia) due to rapid nucleic acid turnover (for example, in leukemia), and especially due to the acute lysis of tumor cells (for example, in successful chemotherapy). In this situation, uric acid crystallization in the kidney can cause kidney failure. Uric acid concentration may also rise in the blood because of diminished renal excretion, the situation in most patients with gout; in renal disease; under conditions such as lactic acidemia or ketoacidosis, in which organic acids such as lactic acid compete with uric acid for tubular reabsorption; and in response to certain diuretics. Hyperuricemia also results from overproduction of purines through the de novo pathway, as occurs in Lesch-Nyhan disease, which is due to a genetically determined deficiency in the HPRT enzyme. Other symptoms of Lesch-Nyhan disease include mental retardation, neurologic abnormalities, and bizarre and self-injurious behavior. Chronic hyperuricemia of any cause may lead to gout, with its painful, disfiguring arthritis and potentially fatal renal failure.

Hyperuricemia that results from diminished renal excretion, but not from renal failure, can be treated with drugs that cause increased excretion of uric acid. However, these drugs are contraindicated in hyperuricemia resulting from overproduction of purines or from nucleic acid breakdown. Treatment is instead by pharmacologic inhibition of xanthine oxidase. *See* GOUT; KIDNEY DISORDERS; METABOLIC DISORDERS; METABOLISM; PROTEIN METABOLISM.

William Nyhan

Bibliography. H. Jinnah, in C. R. Scriver et al. (eds.), *The Metabolic and Molecular Bases of Inherited Disease*, p. 2573, McGraw-Hill, New York, 2001; W. L. Nyhan and P. Ozand, *Atlas of Metabolic Diseases*, p. 376, Chapman & Hall-Arnold, London, 1998.

Uridine diphosphoglucose (UDPG)

A compound in which α-glucopyranose is esterified, at carbon atom 1, with the terminal phosphate group of uridine-5′-pyrophosphate (that is, uridine diphosphate, UDP). On very mild acid hydrolysis, glucose and UDP are liberated. Uridine diphosphoglucose occurs in animal, plant, and microbial cells and is synthesized enzymatically from uridine triphosphate (UTP) and α-glucose-1-phosphate by UDP-glucose pyrophosphorylase, as shown in reaction (1).

$$\text{Glucose-1-phosphate} + \text{UTP} \rightarrow \text{UDP-glucose} + \text{pyrophosphate} \quad (1)$$

This compound functions as a key in the transformation of glucose to other sugars. It can, for example, be epimerized to UDP-galactose, oxidized to UDP-glucuronic acid (which in turn can be decarboxylated to UDP-xylose and UDP-arabinose), or reduced to UDP-rhamnose. All of these compounds are then glycosyl donors utilized in the synthesis of simple compounds (such as morphine glucuronide, sucrose, or lactose) or complex polysaccharides (such as glycogen and cellulose). The general reaction is (2).

$$\text{UDP-X} + \text{ROH} \rightarrow \text{R-X} + \text{UDP} \quad (2)$$

The interconversion of UDP-glucose and UDP-galactose is also important in the utilization of free galactose by cells. Galactose is first phosphorylated to form galactose-1-phosphate. It then reacts with UDP-glucose to yield UDP-galactose and glucose-1-phosphate, and the UDP-galactose is finally epimerized to UDP-glucose. Since the UDP-glucose is regenerated in the last step, it is a cofactor in the sequence and is not used up. The overall reaction is (3).

$$\text{Galactose} + \text{ATP} \rightarrow \text{glucose-1-phosphate} + \text{ADP} \quad (3)$$

UDP-glucose was the earliest discovered of a class of nucleoside diphosphosugar compounds. Other transformations and syntheses occur with glucose attached to other nucleotides. Also, several sugars other than glucose may be initially attached to nucleotides.

For example, cytidine diphosphoglucose (CDP) and thymidine diphosphoglucose (TDP) can be

transformed to CDP-3,6-dideoxyhexoses or TDP-L-rhamnose, which then function as glycosyl donors. Adenosine diphosphoglucose and guanosine diphosphoglucose are primary glycosyl donors; the former is the intermediate in starch biosynthesis in plants, playing a role analogous to that of UDP-glucose in glucogen synthesis in animals.

UDP-acetylglucosamine is formed from acetylglucosamine-1 phosphate and UTP, and guanosine diphosphomannose is formed from mannose-1 phosphate and GTP. The former can be epimerized to UDP-acetylgalactosamine or transformed to complex peptide derivatives of UDP-acetylmuramic acid (compounds involved in the synthesis of bacterial cell walls and the mechanism of action of penicillin). GDP-mannose can also undergo a number of transformations (for example, to GDP-mannuronic acid) or can function directly as a glycosyl donor.

In general, then, UDP-glucose is a prominent member of a family of compounds composed of sugars activated as derivatives of nucleoside diphosphates. In these active forms they can be used directly as glycosyl donors, or they may be transformed to other more complex sugars before their utilization for the biosynthesis of both simple saccharides and complex polysaccharides. *See* CARBOHYDRATE METABOLISM; NUCLEIC ACID.

Jack L. Strominger

Bibliography. S. J. Baum, *Introduction to Organic and Biological Chemistry*, 1993; D. M. Greenberg (ed.), *Metabolic Pathways*, vol. 1, 3d ed., 1967; L. Stryer, *Biochemistry*, 4th ed., 1995.

Urinalysis

Laboratory examination of urine. Urine is a filtrate of the blood and is produced in the kidneys. It is a reflection of the metabolic activity of the body; conditions that affect the normal homeostatic mechanisms are often revealed by a careful analysis of the composition of the urine. *See* KIDNEY; KIDNEY DISORDERS.

Modern routine urinalysis can be divided into two basic procedures: macroscopic and chemical examination, and microscopic analysis.

Macroscopic and chemical examination. This has two separate segments, with most of the chemical examination carried out by using dipstick testing.

Macroscopic examination. This segment includes noting the color and clarity of the urine. Normal urine is pale yellow or straw colored and is usually transparent or clear; abnormal urine may vary greatly in color and may show varying degrees of cloudiness. In addition, any abnormal odor of the urine is noted. Certain drugs and infectious agents make products with distinct odors; for example, diabetic urine often has a "fruity" odor. *See* DIABETES.

The specific gravity of urine, that is, the ratio of the weight of a volume of urine to that of the same volume of water, is measured routinely. Since urine contains the products of renal metabolism, such as salt and nitrogenous wastes, it is heavier than water. Specific gravity is an indicator of the kidney's ability to concentrate or dilute the urine and thus of renal tubular function. When the kidneys are functioning poorly, the specific gravity may be fixed at 1.010 (normal range is 1.003–1.032) because the renal tubules are unable to adequately concentrate or dilute the urine.

TABLE 1. Values for a normal urine sample as measured by dipstick testing

Test or constituent	Presence or concentration
Volume	1000–1500 ml/day
Appearance	Clear; straw-colored
pH	Range 4.5–8; usually acid; diet-dependent
Protein	2–8 mg/dl per urine sample [less than 150 (mg/dl)/day]
Homoglobin (blood)	Not normally detectable
Glucose (sugar)	0–20 mg/dl
Ketone bodies	0–2 mg/dl [less than 20 (mg/dl)/day]
Bilirubin	0.02 mg/dl
Urobilinogen	0.5–4 mg/day
Nitrite (bacterial metabolic product)	Not normally detectable
Leukocyte esterase (product of neutrophils)	Not normally detectable
Specific gravity	1.003–1.032

Chemical testing. Most routine chemical urinalysis is now carried out by dipstick testing, which involves the use of plastic strips, or dipsticks, bearing pads embedded with chemical reactants and color indicators. When the strips are dipped into urine, the chemicals in the pads react with it and change color. The reaction of each pad represents a separate chemical test for a specific product in the urine. The detection of chemical abnormalities in the urine often plays a vital role in diagnosis and therapy since it provides a simple and noninvasive means of assessing renal and general metabolic function (**Table 1**). Dipstick testing includes the following categories:

1. Protein. Protein, predominantly albumin, is normally present in urine but only in small amounts (usually less than 150 mg per day). Increased protein excretion (proteinuria) is sometimes associated with heart or kidney failure or infectious diseases.

2. Glucose. Reducing sugars, of which glucose is the most important, are generally not found in normal urine. Glucose in the urine (glycosuria) is usually associated with diabetes mellitus.

3. Ketone bodies. Ketone bodies, in the form of acetone, acetoacetic acid, and beta-hydroxybutyric acid, are present in normal urine in very small quantities. Elevation of these substances in the blood (ketosis) or in the urine (ketonuria) is usually related to abnormalities in carbohydrate metabolism, for example, diabetes mellitus.

4. Blood. Hemoglobin, a component of red blood cells, is normally found in minute quantities in the urine. Excessive hemoglobin in the urine (hemoglobinuria) is a pathologic finding, since it means that blood cells are leaking out of the bloodstream.

5. Bile. Small amounts of bilirubin normally can be found in urine. Larger amounts occur in jaundice

that is produced by biliary tract obstruction. *See* PROTEIN.

Microscopic examination. The urine normally contains a wide variety of formed elements that can be identified by using a light microscope. Together these elements form the urinary sediment.

Cells. Cells in the urine originate in the bloodstream or in the epithelium lining the urinary tract. The main types of epithelial cells are renal tubular cells, transitional (urothelial) cells, and squamous cells. All three types occur in relatively small numbers in the normal sediment. In disease states that affect the nephron, increased numbers of renal tubular cells may be present in the urine. Transitional cells may show an increase after urethral catheterization or prostate surgery, as well as in the presence of bladder infections or bladder cancer. Increased numbers of squamous cells in the urine may be a urinary contaminant from the vaginal epithelium.

Blood cells occur normally in the urine in small numbers, and consist mainly of polymorphonuclear neutrophils (a type of granular leukocyte) and red blood cells. A high level of polymorphonuclear neutrophils may indicate an infection of the urinary tract. Increased red blood cells may indicate a "silent" tumor of the kidney or bladder or, when accompanied by casts, may indicate a disease specifically affecting the renal glomeruli.

Casts. These proteinaceous products of the kidney are of major importance when present in increased numbers or seen in abnormal forms, because they usually indicate intrinsic renal disease. They are formed only in the kidney—in the lumen of the distal nephron—and consist mainly of Tamm-Horsfall protein, which is unique to the kidney. Casts are cylindrical and are named on the basis of their microscopic appearance and the cells they contain (**Table 2**).

TABLE 2. Types and characteristics of normal and abnormal urinary casts

Type of cast	Characteristics
Associated with normal urine	
Hyaline	Smooth surface; low refractive index; hard to see without staining
Granular	Fine granules scattered throughout
Associated with abnormal urine	
Granular	Coarse granules scattered throughout
Waxy	High refractive index; easy to see; surface looks waxy; notched sides; ends squared off (look broken off)
Fatty	Cast surface covered with high refractile lipid globules of varying size
Pigment	Basically a hyaline cast that is colored because of absorption of an abnormal pigment such as hemoglobin or melanin
Red blood cell	Surface covered with red blood cells
White blood cell	Surface covered with polymorphonuclear neutrophils
Renal tubular epithelial	Surface covered with renal tubular epithelial cells
Bacterial/fungal	Surface admixed with polymorphonuclear neutrophils and bacteria or fungi

Microorganisms. Normally, urine is sterile, and the urinary sediment should not contain microorganisms. However, the terminal third of the urethra and the vagina can be inhabited by bacteria and yeast forms, and these may be found in the sediment in small numbers in freshly voided urine. In patients with serious urinary tract infections, microorganisms are usually present in the urine in considerably greater numbers.

Parasites are not normally found in the urine. They occur rarely in the United States but are present more frequently in other parts of the world. *Schistosoma hematobium* is the only human parasite indigenous to the urinary tract; other parasites can appear in the urine as contaminants from the vagina or the perirectal area. Lice and mites from an individual can fall into urine during voiding.

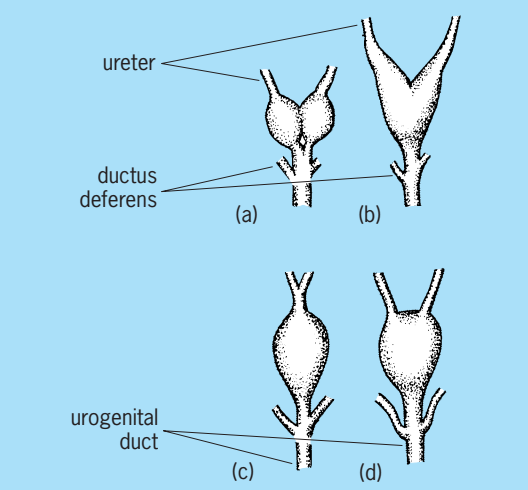
Other elements. Mucus is frequently found in urine sediment and has no known pathologic significance. A wide variety of crystals appear in the urine; their presence may be normal or may indicate an abnormal state. In addition, crystalline forms of certain medications, such as sulfa drugs, are often seen in urine sediment. Crystal identification is performed with the aid of polarized microscopy, because many crystals in the urine are birefringent. *See* POLARIZED LIGHT MICROSCOPE.

Additional tests. Other nonroutine tests may be performed on urine. Renal function can be tested, and renal clearance can be measured. Also, tests can be carried out for the presence of poisons, drugs, and abnormal metabolites, and for evidence of pregnancy. *See* CLINICAL PATHOLOGY. Meryl H. Haber

Bibliography. J. A. Freeman and M. F. Beeler, *Laboratory Medicine*: *Urinalysis and Medical Microscopy,* 2d ed., 1983; M. H. Haber, *The Urinary Sediment: A Textbook Atlas*, 1981; T. A. Stamey and R. W. Kindrachuk, *Urinary Sediment and Urinalysis: A Practical Guide for the Health Science Professional,* 1985; S. K. Strasinger, *Urinalysis and Body Fluids*, 4th ed., 2001.

Urinary bladder

A distensible, muscular sac in most vertebrates which serves as a reservoir for urine. Snakes, crocodilians, birds (with the exception of the ostrich), most lizards, and a few fish lack a urinary bladder. In these organisms, urine empties directly into the cloaca. The development of the urinary system is intimately associated with the development of the reproductive system. Three general types of urinary bladder are recognized among the vertebrates: tubal, cloacal, and allantoic. *See* URINE.



Types of tubal bladders. (*a*) Duplex. (*b*) Bilobed. (*c*) Simplex, with ureters united. (*d*) Simplex, with separate ureters. (*After H. E. Walter and L. P. Sayles, Biology of the Vertebrates, 3d ed., Macmillan, 1949*)

Tubal bladders. Most fish possess tubal bladders, that is, enlargements of the mesonephric ducts (see **illus.**). A single bladder, simplex type, results when two excretory ducts unite to form a single expanded structure. In the cod (*Gadus*) two separate bladders, duplex type, occur near the terminus of each excretory duct. These structures unite posteriorly to form a single passageway. A third type of tubal bladder is the bilobed bladder which occurs in gars as the result of fusion of the left and right urinary ducts.

Cloacal bladders. The cloacal bladder is found in monotremes, amphibians, and some dipnoans. There is no direct connection between the excretory ducts and this type of bladder. The bladder is an outpouching or diverticulum of the cloacal wall. The cloacal opening is closed by a sphincter muscle and the urine which seeps into the cloaca from the excretory ducts is forced into the bilobed bladder.

Allantoic bladders. This type of bladder is derived from the ventral wall of the cloaca and possibly the allantoic diverticulum. The role of the allantois in the formation of this type of bladder, which is found in most mammals, the turtles, and those lizards which have a bladder, is questioned by some embryologists. *See* ALLANTOIS.

The mammalian bladder is lined with a special epithelium composed of transitional cells. The muscular layer is composed of vertical, horizontal, and oblique fibers. The bladder drains through the urethra, the opening being controlled by a sphincter. Innervation is by the hypogastric sympathetic plexus and partly by parasympathetic fibers from the second and third sacral nerves. Stimulation of the parasympathetic causes the bladder muscle to contract and relaxes the internal sphincter. Micturition is a reflex act which is initiated voluntarily except in children. *See* PARASYMPATHETIC NERVOUS SYSTEM; SYMPATHETIC NERVOUS SYSTEM; URINARY SYSTEM. Charles B. Curtin

Urinary system

The system consisting of the kidneys, urinary ducts, and bladder. These structures will be discussed in this article in terms of their comparative anatomy, embryology, and physiology.

Comparative Anatomy

Similarities are not particularly evident among the many and varied types of excretory organs found among vertebrates. The variations that are encountered are undoubtedly related to problems with which vertebrates have had to cope in adapting to different environmental conditions.

Archinephros. It is now generally believed that the primitive vertebrate ancestor possessed an excretory organ referred to as an archinephros or holonephros (**Fig. 1**). This probably consisted of a pair of dorsally located ducts extending the length of the body cavity. Each duct was joined by a series of segmentally

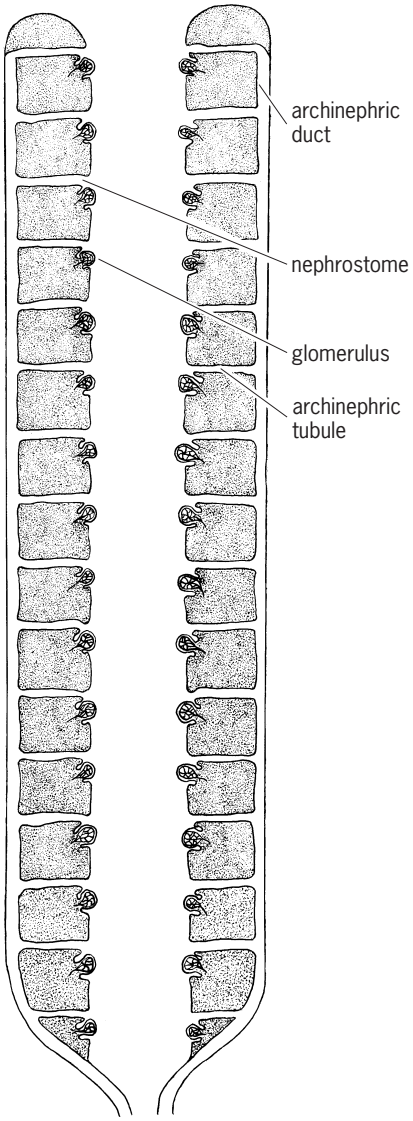


Fig. 1. Hypothetical structure of archinephros. (*After C. K. Weichert, Elements of Chordate Anatomy, 3d ed., McGraw-Hill, 1967*)

arranged tubules, one pair to each segment. The other end of each tubule opened into the body cavity by a ciliated, funnel-shaped aperture. Close to each opening was a small knot of arterial blood vessels called an external glomerulus. From this type of kidney with its archinephric duct the various kidneys of forms living today may originally have been derived. The larval form of the hagfish and the larvae of certain amphibians, the caecilians, are present-day vertebrates possessing kidneys of this type.

Anamniote kidneys. The anterior portion of the archinephric kidney persists only in the adult stage of the hagfish and of certain teleost fishes in which it is called the head kidney, or pronephros. It appears in the embryos of most vertebrates as a transitory structure that usually degenerates soon after it has formed. The remainder of the kidney posterior to the pronephros is known as the opisthonephros.

Pronephros. The importance of the pronephros lies mainly in the part it plays during development in forming the archinephric duct which persists even though the pronephros appears only as a transient structure. In some larval forms the pronephros may be important in getting rid of wastes at a time when the opisthonephros is being formed. Even in the hagfish, the head kidney has become modified from the primitive condition. Here the openings of the tubules connect with the pericardial cavity, and the fluid drained by the tubules passes into a nearby vein instead of entering the archinephric duct. Pronephros and opisthonephros become completely separated by degeneration of the portion between them.

Opisthonephros. Because the pronephros is usually a transitory structure, the opisthonephros is the more important of the two. It serves as the adult kidney in lampreys, most fishes, and amphibians. The opisthonephros differs from the pronephros in several respects. A main distinction is that the segmental arrangement of the kidney tubules is lost and many tubules may lie within the confines of a single segment. Furthermore the connection of the opisthonephric tubules with the body cavity is usually lost, and renal corpuscles with internal glomeruli are typically present. These are small knots of arterial vessels, each surrounded by a double-walled cup called Bowman's capsule; the two together form a renal corpuscle (**Fig. 2**). The internal, or visceral, layer of Bowman's capsule is actually very complexly arranged, folded, and interdigitated to follow the configurations of the individual glomerular capillaries which exhibit extensive anastomoses. The cells, called podocytes, are not squamous in type and do not furnish a continuous covering of the capillaries.

In some forms the anterior tubules of the opisthonephros lie in the same segments as posterior pronephric tubules. This indicates the transitional nature of the two. A typical opisthonephric tubule consists of a narrow neck adjacent to the renal corpuscle, followed in turn by secretory and collecting portions. The collecting portion joins the archinephric duct. The ends of several collecting tubules may unite to form a common duct which either opens into the archinephric duct or establishes an independent connection with the cloaca. Several such accessory ducts may be present.

Cyclostomes. The opisthonephros of the adult hagfish differs basically from the original archinephros only in the loss of peritoneal connections of the posterior tubules. In the adult lamprey the opisthonephros on each side consists of a long, strap-shaped body without peritoneal connections. The kidneys lie on either side of the middorsal line from which each is suspended by a mesenterylike membrane. The archinephric duct lies along the free edge of the kidney. The ducts of the two sides unite posteriorly to open into a urogenital sinus which leads to the outside through an aperture at the tip of a small urogenital papilla. Two slitlike genital pores connect the urogenital sinus with the coelom. The condition is similar in both sexes. Eggs or spermatozoa leave the body cavity by way of the genital pores, the urogenital sinus, and the urogenital aperture. Only here are the reproductive and urinary systems associated.

Fishes. There is much variation in shape of the opisthonephric kidneys of fishes, but they are fundamentally similar in structure. In some they extend the length of the coelom; in others they are short and may show various degrees of fusion. Peritoneal funnels rarely occur. Some marine teleosts lack glomeruli and thus possess aglomerular kidneys. In elasmobranchs the anterior ends of the kidneys of the male have been appropriated by the reproductive system. Modified kidney tubules, called efferent ductules, connect each testis with the archinephric duct which lies on the ventral surface of the kidney and serves as a ductus deferens for sperm transport. Accessory urinary ducts are usually present. In teleost fishes there is no connection between the testes and the opisthonephric kidneys. The posterior ends of the archinephric ducts of female fishes

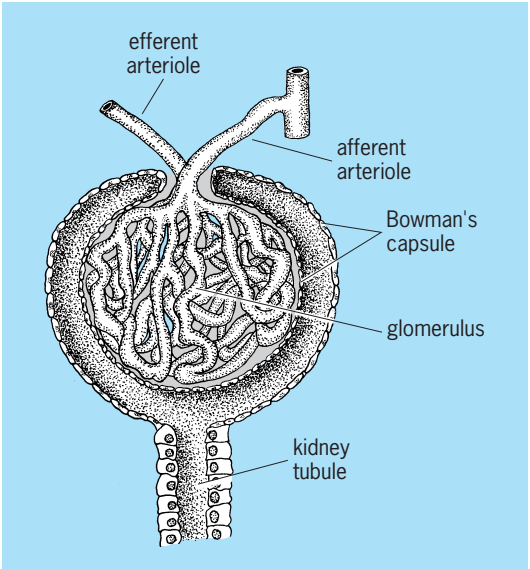


Fig. 2. Renal corpuscle. (*After C. K. Weichert, Anatomy of the Chordates, 4th ed., McGraw-Hill, 1970*)

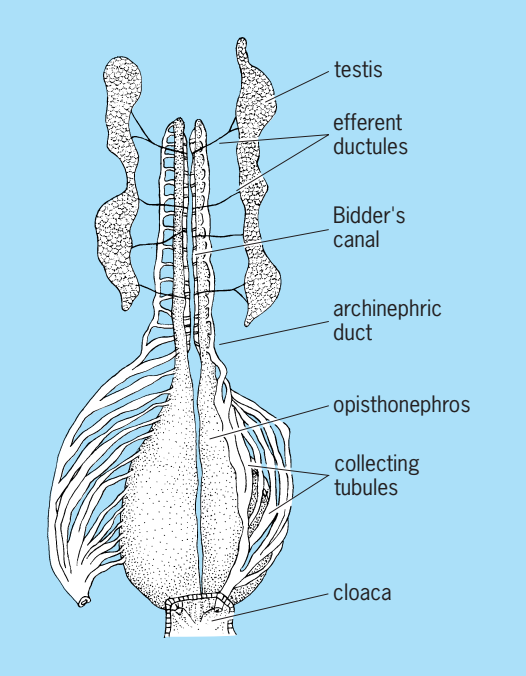


Fig. 3. Urogenital organs of male salamander, ventral view. The collecting ducts on the left are shown detached from the cloaca and spread out for clarity. (*After C. K. Weichert, Elements of Chordate Anatomy, 3d ed., McGraw-Hill, 1967*)

enter a common urinary sinus inside a small urinary papilla. The latter enters the cloaca in elasmobranchs and dipnoans, but in most other fishes it opens directly to the outside, a cloaca being absent.

Amphibians. In common with other amphibians, adult caecilians possess an opisthonephros. The kidneys of the tailed amphibians are much like those of elasmobranch fishes, the anterior ends in males being concerned with genital rather than urinary functions (**Fig. 3**). The archinephric duct courses along the lateral edge of the kidney a short distance outside the kidney proper. Numerous collecting ducts or tubules leave the kidney at intervals to join the duct. The two ducts in both sexes open separately into the cloaca. In frogs and toads the opisthonephric kidneys lie toward the posterior part of the abdominal cavity. A yellowish adrenal gland is located along the ventral surface of each. The kidneys of females are not related to the reproductive system but in males an intimate connection exists. Modified tubules, or efferent ductules, from the testes connect with kidney tubules which lead to the archinephric duct. This duct is located within the kidney along its lateral margin. Peritoneal funnels are present on the ventral sides of the kidneys in some frogs but they connect with veins rather than the kidney tubules. A thin-walled urinary bladder opens into the amphibian cloaca. It has no connection with the archinephric ducts.

Amniote kidneys. In reptiles, birds, and mammals three types of kidneys are usually recognized: the pronephros, mesonephros, and metanephros. These appear in succession during embryonic development, but only the metanephros persists in the adult. Mesonephros and metanephros actually represent different levels of the opisthonephros of lower forms, the metanephros being equivalent to the posterior portion.

The anteriorly located pronephros appears during very early development, but it soon degenerates and the more posterior mesonephros then develops. The duct of the pronephros, however, persists to become the duct of the mesonephros. This is actually the same as the archinephric duct but is usually referred to as the Wolffian duct.

The mesonephros persists for a time and then degenerates. In the meantime the metanephros has begun to develop from the region posterior to the mesonephros. A few mesonephric tubules and the Wolffian duct persist to contribute to the reproductive system of the male or to remain as vestigial structures.

Mesonephros. The tubules of the mesonephros develop in the same manner as opisthonephric tubules. Some of the anterior tubules may even form peritoneal connections. In some forms the mesonephric kidneys become voluminous structures; in others they amount to very little. In reptiles, spiny anteaters, and marsupials the mesonephros may even persist for a time after birth.

Metanephros. A metanephric kidney develops on each side posterior to the mesonephros. It is composed of essentially the same parts as the mesonephros and contains renal corpuscles, secretory tubules, and collecting tubules. No peritoneal connections are present. Each kidney has a twofold origin. An outgrowth from the posterior end of the Wolffian duct grows forward into the tissue posterior to that from which the mesonephros was derived and which is called the metanephric blastema. The outgrowth is destined to form the ureter and the collecting portion of the kidney. It branches and rebranches many times to form ultimately large numbers of fine collecting tubules. At least in mammals at the point where the outgrowth undergoes its primary divisions an expanded region forms the pelvis of the kidney.

From the blastema adjacent to the collecting tubules arise secretory tubules. Each tubule grows and becomes S-shaped. One end establishes connection with a collecting tubule; the other expands and becomes invaded by a vascular glomerular tuft so that a typical renal corpuscle is formed. Each tubule differentiates into several regions (**Fig. 4**).

Reptiles. The metanephric kidneys of reptiles lie in the posterior part of the abdominal cavity, usually in the pelvic region. They are small, compact, and often markedly lobulated. The posterior portion on each side is somewhat narrower. In some lizards the hind parts may even fuse. The degree of symmetry varies, being most divergent in snakes and limbless lizards which have notably long, narrow, lobulated kidneys in correlation with the shape of the body. One kidney may be entirely behind the other.

Snakes and crocodilians lack a urinary bladder, but most lizards and turtles have well-developed bladders which open into the cloaca. Except in turtles

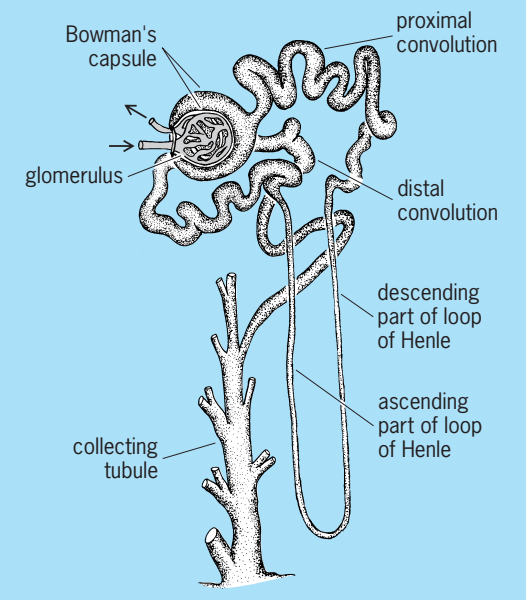


Fig. 4. Mammalian metanephric tubule, showing the renal corpuscle and secretory and collecting portions. (*After C. K. Weichert, Elements of Chordate Anatomy, 3d ed., McGraw-Hill, 1967*)

the ureters open independently into the cloaca. In turtles they connect to the bladder. Some turtles possess a pair of accessory urinary bladders which open into the cloaca. They may be used as accessory organs of respiration. In females they are reported to be filled with water which is used to soften the soil when a nest is being prepared.

Birds. The kidneys of birds are situated in the pelvic region of the body cavity; their posterior ends are usually joined. They are lobulated structures with short ureters which open independently into the cloaca.

Except for the ostrich, birds lack urinary bladders. Urinary wastes, chiefly in the form of semisolid uric acid, are eliminated through the cloaca along with the feces.

Mammals. The rather typical mammalian metanephric kidney (**Fig. 5**) is a compact, bean-shaped organ attached to the dorsal body wall outside the peritoneum. The ureter leaves the medial side at a depression, the hilum. At this point a renal vein also leaves the kidney and a renal artery and nerves enter it. The kidney is surrounded by a capsule of connective tissue under which lies the cortex. The renal corpuscles and the greater part of the secretory tubules lie entirely in the cortex. The portion of the kidney surrounded by the cortex is the medulla. It is partly composed of large areas, the renal pyramids. The outer borders of the pyramids are divided into smaller units called lobules. The collecting tubules lie within the pyramids but may extend well up into the cortex. The inner portion of each pyramid, in the form of a blunt papilla, projects into an outpocketing of the pelvis known as a minor calyx. Several minor calyces join together to enter major calyces which in turn open into the renal pelvis. The pelvis leads to the ureter which empties into the bladder, except in monotremes in which it enters the urethra. Urine, which is stored temporarily in the bladder, passes to the outside through the urethra. In males the urethra opens at the tip of the penis; in females the condition varies, for in some, as the rat and mouse, the urethra opens independently to the outside, passing through the clitoris. It usually, however, enters a vestibule which is the terminal part of the genital tract. The kidneys of mammals are markedly lobulated in the embryo, and in many forms this condition is retained throughout life. *See* KIDNEY. Charles K. Weichert

Comparative Embryology

The kidney, or nephros, of vertebrates is made up of many individual structural and functional units known as nephrons. The nephrons are derived from that portion of the embryonic mesoderm designated as the intermediate mesoderm. Ideally, this material becomes segmented, with each segment being termed a nephrotome (**Fig. 6**). Any given nephrotome contains a coelomic chamber, the nephrocoele, which opens to the adjacent body cavity via a peritoneal funnel.

The conversion of nephrotome to nephron involves the following events. There arises from the dorsolateral wall of the nephrotome a tubular outgrowth, the principal tubule, which communicates with the nephrocoele via a nephrostome; the medial wall of the nephrotome thins, flattens, and bulges inwardly coincidently with its invasion by arterial capillaries derived from the nearby dorsal aorta. The skein of capillaries makes up a glomerulus, and the wall of the nephrotome investing the glomerulus is known as a renal, or Bowman's, capsule (**Fig. 7**).

The original vertebrate nephros presumably consisted of similar nephrons throughout the length of the organ with each opening independently to the

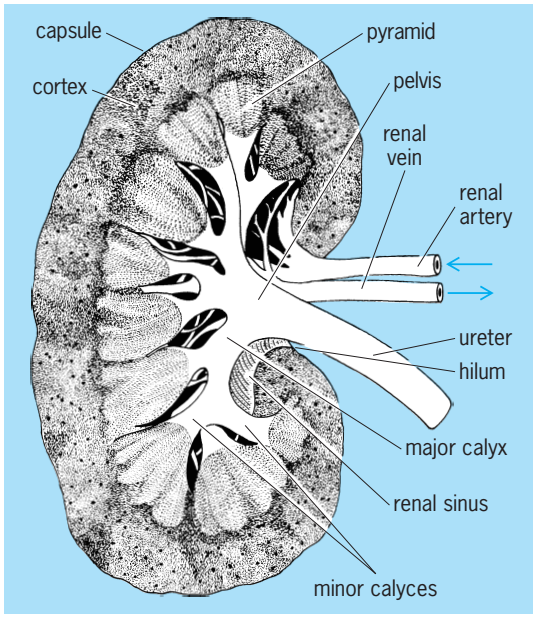


Fig. 5. Sagittal section of metanephric kidney of a human (semidiagrammatic). (*After C. K. Weichert, Elements of Chordate Anatomy, 3d ed., McGraw-Hill, 1967*)

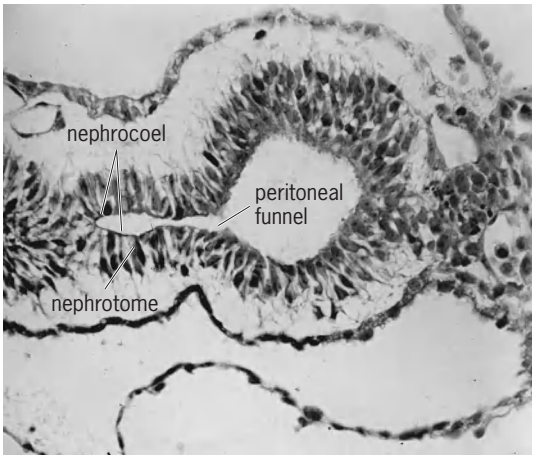


Fig. 6. Nephrotome of human embryo.

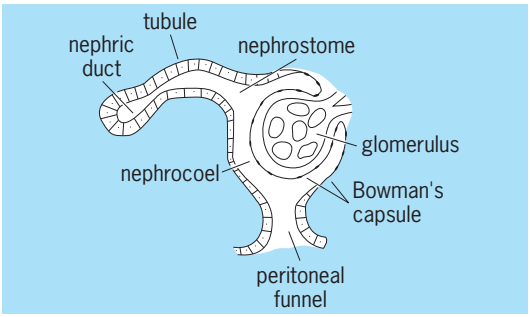


Fig. 7. Basic design of a nephron.

exterior. However, this arrangement is only hypothetical, for in all known vertebrates the nephrons empty into a common drainage, or nephric, duct which passes back alongside the nephros to terminate in the cloaca, the chamber which also receives the digestive tract. The situation is also complicated in present-day vertebrates by four variables. First, in the embryos of higher vertebrates, typical hollow nephrotomes are seldom formed; instead, nephrons differentiate without segmental arrangement within a continuous cord, the nephrogenic cord, of intermediate mesoderm. Second, as the nephros develops embryonically, its entire length does not appear at one time; instead, the nephrons appear in sequence from front to rear and the first-formed anterior ones tend to disappear before the posterior ones arise. Third, the nephrons become progressively more complex from anterior to posterior. Fourth, the manner of establishment of the nephric duct is not consistent in all vertebrates. The results of these four variables follow.

Nephros of fishes and amphibians. Embryonic development of the nephros is inaugurated within the most anterior reaches of the mesomere. This intermediate mesoderm becomes segmented into nephrotomes from each of which a nephron forms in the general manner already described. The number of these first formed nephrons varies with the species, but is always relatively small, usually three to five. These nephrons, because of their anterior position and because they are the first to appear, form the head kidney, or pronephros. Accordingly, the nephrons themselves are termed pronephric tubules, or pronephrons. As the first and most anterior pronephric tubule arises, it first extends dorsolaterally and then turns backward to join the one forming immediately behind. This one in turn joins the one behind it, and so on, thus producing a common drainage duct designated the pronephric duct. The pronephric duct, once initiated, extends itself backward along the still undifferentiated intermediate mesoderm until it joins the cloaca (**Fig. 8***a*).

With the exception of the hagfishes and some bony fishes in which it persists throughout life, the pronephros has a temporary existence. In the sharks, for instance, it is present only during early embryonic stages and has little or no functional significance, but in those forms having an active free-living larval stage, such as amphibians and lampreys, it persists in the larva as a functional organ and the individual pronephrons possess ciliated peritoneal funnels opening to the adjacent coelom. This provides for the uptake of certain materials directly from the coelom as well as from the bloodstream. Direct demonstrations of the functional capacities of pronephric tubules have come from a variety of tests. For example, the pronephrons of the larval lamprey will take up quantities of coloidal carbon which may be injected into the coelom; and through their ability to accumulate the dye phenol red, the pronephrons of frog larvae and certain bony fishes have also revealed their functional capabilities.

Whatever the length of its existence, the pronephros is supplemented and succeeded by a second generation of nephrons derived from the remainder of the intermediate mesoderm. Although

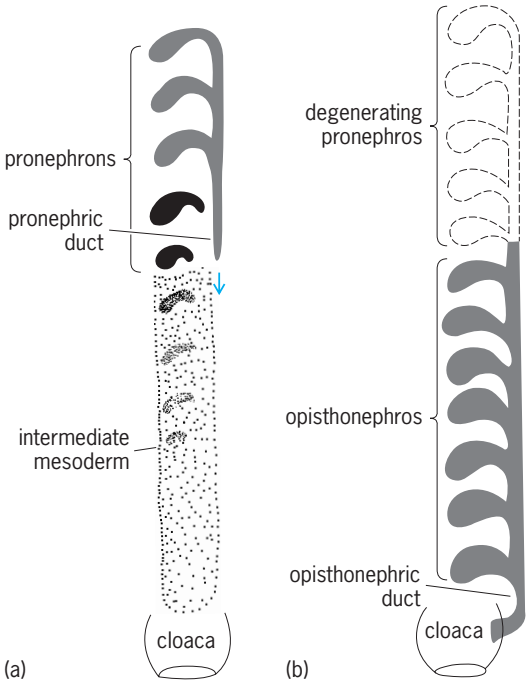


Fig. 8. Diagram of (*a*) developing pronephros and (*b*) opisthonephros.

they arise in basically the same manner and exhibit the same fundamental structure, these later-formed nephrons tend to be longer and more complex in their makeup, and ordinarily lack peritoneal funnels. Unlike their pronephric forerunners these nephrons fail to establish their own drainage duct; they join the already existing pronephric duct. Eventually, as noted, the earlier-formed pronephros disappears, leaving this later generation of nephrons to constitute the final, definitive kidney. This organ, distinctive to fishes and amphibians, is the opisthonephros, or back kidney, and the former pronephric duct which its nephrons have taken over is the opisthonephric duct (Fig. 8*b*).

Nephros of reptiles, birds, and mammals. The embryonic initiation of the nephros of higher vertebrates is customarily described as involving the establishment of a pronephros and pronephric duct as in fishes and amphibians. Although this may be true for reptiles and birds, in mammals pronephric tubules rarely if ever appear. In humans, for example, that level of intermediate mesoderm equivalent to the pronephros never gets beyond the point of conversion to a few rudimentary nephrotomes. Because definitive pronephric tubules are never provided, the pronephric duct must arise in some fashion other than by junction of the ends of pronephrons. The original nephric duct is initiated as a solid rod which splits off the dorsolateral side of the nephrogenic cord (**Fig. 9**). Once established in this fashion, the solid duct then frees itself from the parent material and as a tapering rod extends itself backward by independent terminal growth, ultimately contacting and fusing with the wall of the cloaca. The solid rod gradually hollows out and within a few days after its original establishment becomes tubular throughout.

Wolffian duct. As the nephric duct, or Wolffian duct, is being formed the first of two generations of nephrons appears. The first generation consists of a series of nephrons derived from the middle level

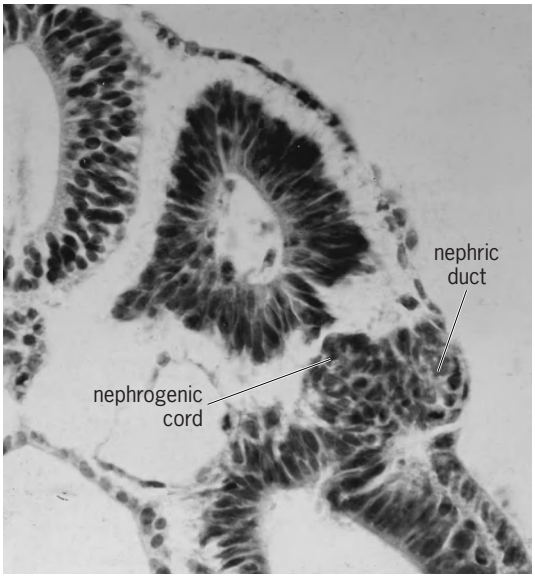


Fig. 9. Origin of nephric duct in human embryo.

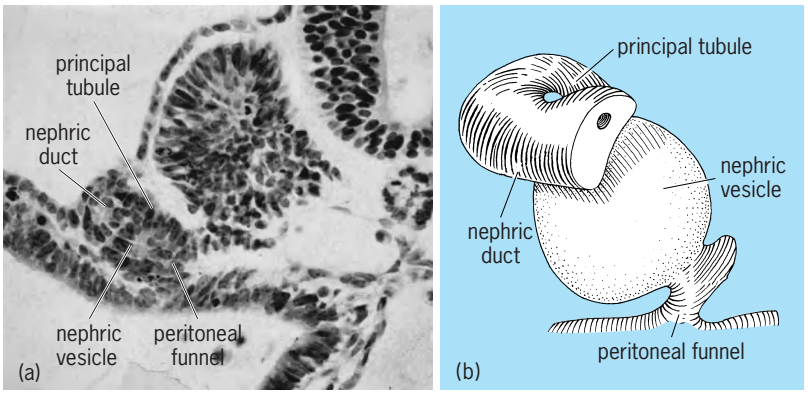


Fig. 10. Early human mesonephron. (*a*) Section. (*b*) Reconstruction.

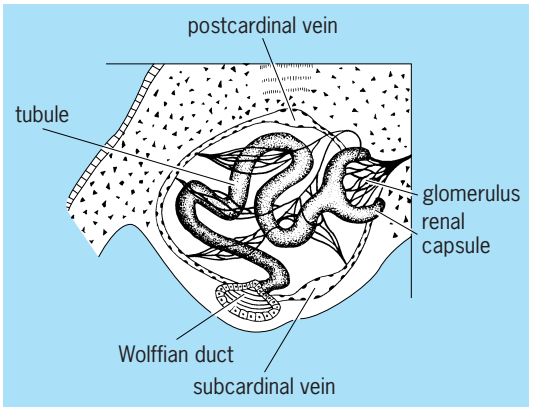


Fig. 11. Fully differentiated human mesonephron. (*After T. W. Torrey, Morphogenesis of the Vertebrates, 2d ed., John Wiley and Sons, 1968*)

of the intermediate mesoderm. The details of their manner of development and final form vary from one category of animals to another, but in general they conform to the pattern described above. The human embryo may be used as a specific illustration. Briefly, the nephrogenic cord paralleling the growing Wolffian duct provides an increasing number of serially arranged spherical bodies known as nephric vesicles (**Fig. 10**). These are solid at first, but become hollow, and as they do so each vesicle sends a principal tubule dorsolaterally to join the Wolffian duct. The vesicle proper will provide the capsule surrounding the later-developing glomerulus. Elongation and twisting of the principal tubule and the acquisition of a capillary network complete the nephron (**Fig. 11**).

Wolffian body. These nephrons are known as mesonephric tubules, or mesonephrons, and collectively they make up the mesonephros, or Wolffian body. This kidney is well developed in reptiles and birds, but in mammals exhibits considerable variation. In the rat embryo, for example, it is quite rudimentary and only a dozen or so abortive nephrons arise. At the other extreme is the pig embryo whose mesonephros is large and bulky and involves several hundred long and convoluted nephrons. The human mesonephros lies between these two extremes.

Embryonic function. The variable status of the mesonephros, especially in mammals, raises the interesting question of the functional role it plays in the economy of the embryo. It is reasonable, of course, to infer function from exhibited structure. More convincing, however, are the direct experimental demonstrations of functional capacities of the mesonephrons. To this end, certain techniques employed to assess kidney functions in adult forms have been applied with profit.

The chick embryo has been a common subject for such tests. Solutions of indigo red and trypan blue injected into the vascular system ultimately appear in the mesonephrons. Another approach has been that of cultivating a fragment of mesonephros in a suitable culture medium to which an indicator such as phenol red has been added. The proximal portions of the tubules pick up the indicator and transport it to their lumina; the distal portions resorb water. Still another method has been the direct identification of nitrogenous wastes deposited in the embryonic bladder, the allantois. *See* ALLANTOIS.

Similar experiments on mammals are complicated by the intrauterine location of the embryo and its association with a placenta. Nevertheless, it has been possible either to inject suitable indicators directly into the embryonic body or to introduce them secondarily by transmission via the placenta from the maternal bloodstream. Such tests on embryos of the rabbit, cat, pig, and pouch-young opossum have provided positive demonstrations of the functioning of both the glomerular filter and the tubule proper.

Metanephros. As the mesonephros is attaining its maximum development, a second generation of nephrons is inaugurated. The history of this group is considerably more complicated than the former and runs briefly as follows. A tubular outgrowth from the Wolffian duct, the ureteric diverticulum, appears close to the entrance of the Wolffian duct into the cloaca and pushes itself anteriorly into the undifferentiated intermediate mesoderm behind the mesonephros. The distal end of this diverticulum enlarges, and the intermediate mesoderm coincidently begins to condense around the enlargement. **Figure 12***a* shows that the proximal segment of the original diverticulum is the ureter, or metanephric duct; the expanded distal end of the diverticulum is the primitive renal pelvis; the condensed mesoderm around the pelvis is the metanephric blastema.

Subsequent events pertain primarily to the pelvis and blastema. The former subdivides first to form the two future major calyces (Fig. 12*b*). These divide and subdivide until several generations of branches are produced. The earlier generations come to represent the minor calyces; the later generations become the collecting tubules (Fig. 12*c* and *d*). As the primitive pelvis carries on this program of subdivision, the blastemal tissue subdivides into a corresponding number of masses. Thus there results a nodule of blastema in association with the end of each prospective collecting tubule (**Fig. 13**). Each such nodule or sphere is the forerunner of a metanephric or uriniferous tubule. The solid sphere first becomes a vesicle which, by elongation, is transformed into a tortuous tubule (Fig. 13). The thinner-walled, blind end of the tubule becomes the capsule surrounding the glomerulus forming concomitantly. Coincidentally, the tip of the prospective collecting duct grows out to meet the end of the tubule and the two unite (Fig. 13). The uriniferous tubules, collecting tubules, and calyces are the definitive kidney or metanephros of the late embryo and adult. In terms of gross anatomy, the convoluted portions of the uriniferous tubules, with their glomerular capsules, collectively form the cortex of the kidney. These tubules lead into the collecting system and calyces making up the medulla of the kidney. All drainage ultimately converges upon the renal pelvis, which in turn leads to the ureter.

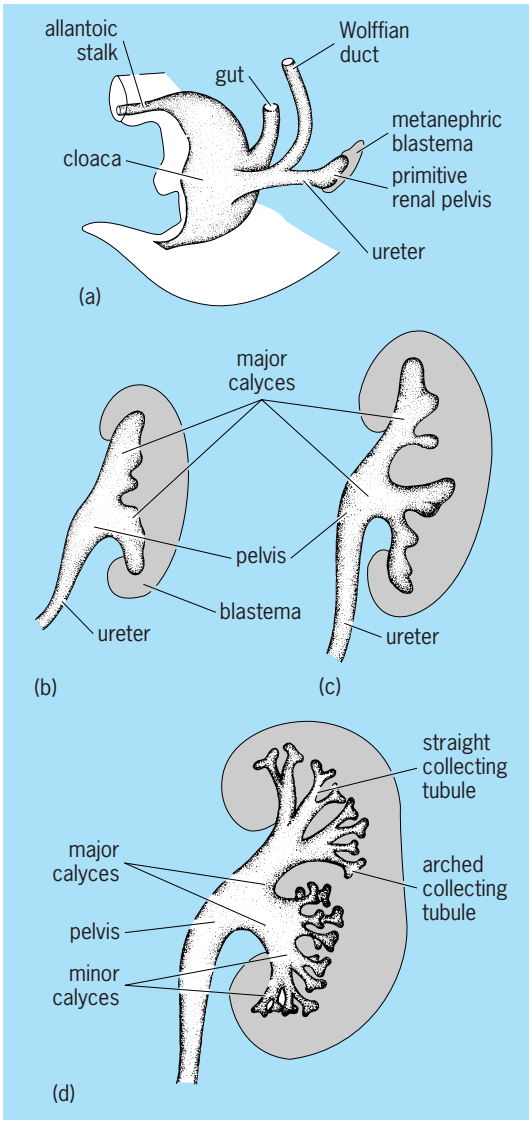


Fig. 12. Stages *a-d* in development of metanephric pelvic and blastema. (*After T. W. Torrey, Morphogenesis of the Vertebrates, 2d ed., John Wiley and Sons, 1968*)

Developmental interdependence. Experimental analyses have revealed important developmental interdependencies within the mesonephric and metanephric

systems. The original nephric duct grows back from the level of its origin to join the cloaca. In the normal course of events it is joined along the way by mesonephrons and thus becomes the Wolffian duct. If the backward extension of the duct alongside the prospective mesonephros-forming area is prevented, little or no development of the mesonephros will occur. This can be accomplished readily in the chick embryo by producing a minute wound or inserting some block in the pathway of the duct. This is interpreted to mean that the duct serves as an inductor of the mesonephric tubules; that is, differentiation of the mesonephrons depends upon some kind of stimulus provided by the nephric duct. Another consequence of such a blockage of the nephric duct is the elimination of the ureter which normally grows from it. This in turn leads to a failure of the blastema to produce metanephric tubules. The ureter is the inductor of these tubules. Development of the nephros is thus revealed as a series of steps, each dependent upon the one before, with the original nephric duct playing the starting role.

During the latter part of embryonic life the mesonephros and metanephros function simultaneously. Gradually, however, the mesonephros regresses and the metanephros assumes full responsibility for excretion. In females, the mesonephros disappears almost entirely; in males, parts of it are incorporated in the reproductive system. As a consequence of straightening and elongation of the fetal body, the kidneys come to be displaced relatively far forward in the body. *See* REPRODUCTIVE SYSTEM.

Urinary bladder. At or near the posterior ends of the nephric ducts there frequently is a reservoir for urine. This is the urinary bladder. Actually there are two basic varieties of bladders in vertebrates. One is found in fishes in which the reservoir is no more than an enlargement of the posterior end of each urinary duct. Frequently the urinary ducts are conjoined and a small bladder is formed by expansion of the common duct. The far more common type of bladder is that exhibited by tetrapods. This is a sac which originates embryonically as an outgrowth from the ventral side of the cloaca. Present in all embryonic life, it is exhibited differentially in adults. All amphibians retain the bladder, but it is lacking in snakes, crocodilians, and a few lizards; birds, also, with one or two exceptions, lack a bladder. It is present in all mammals. Because much of the developmental history of the tetrapod bladder is linked to the history of the cloaca, it will be considered in this conjunction.

Amphibians. The basic pattern is exemplified by the Amphibia. All retain the cloaca as adults and the urinary bladder arises as a diverticulum from the floor of the cloaca (**Fig. 14**). Commonly it is bilobed. There is no direct connection between the excretory ducts and the bladder. Instead, urine first passes into the cloaca and thence into the bladder. Urine is expelled by the intermittent opening of the cloacal orifice and the coincidental contractions of the muscular wall of the bladder.

Reptiles and birds. In reptiles and birds the cloaca is partly subdivided so that the intestine and urogenital ducts open into separate compartments which then join in a common outlet. In the embryo a pouch develops from the floor of the urogenital portion of the cloaca and expands to form a prominent sac known as the allantois. Ultimately the allantois enlarges to extend beyond the confines of the embryo and to fuse broadly with the outermost membrane, the chorion, surrounding the embryo (**Fig. 15**). The combined chorioallantoic membrane ultimately makes broad contact with the membrane lining the inner surface of the shell. This fusion and spreading of the allantois serves to bring an extensive blood supply adjacent to the shell and thus sets up a medium for respiratory exchange. The allantois, therefore, serves not only as a reservoir for urine but also plays a major role in embryonic respiration. *See* FETAL MEMBRANE.

In birds the entire allantois is discarded at hatching and no adult urinary bladder is retained. This is also true for most reptiles although some, notably the turtles and certain lizards, retain the base of the allantois as an adult bladder, and the remainder atrophies.

Mammals. Among mammals, only a few primitive forms retain a cloaca as adults. In all the others it is modified in such a way as to be eliminated.

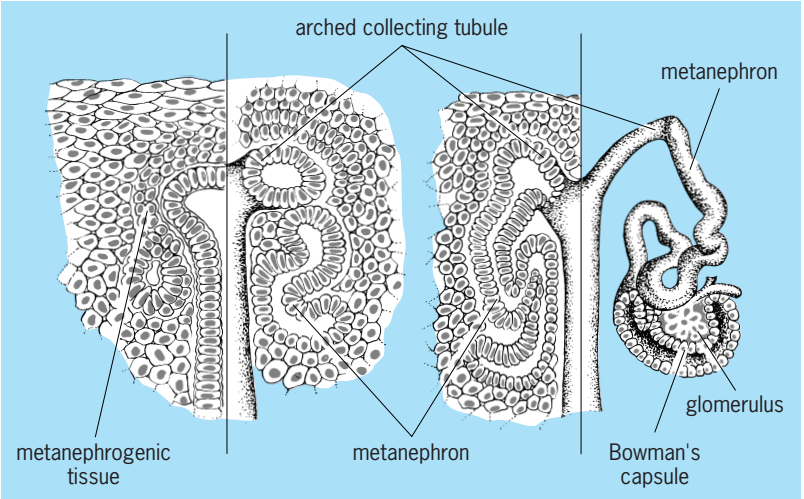


Fig. 13. Differentiating metanephron. (*After T. W. Torrey, Morphogenesis of the Vertebrates, 2d ed., John Wiley and Sons, 1968*)

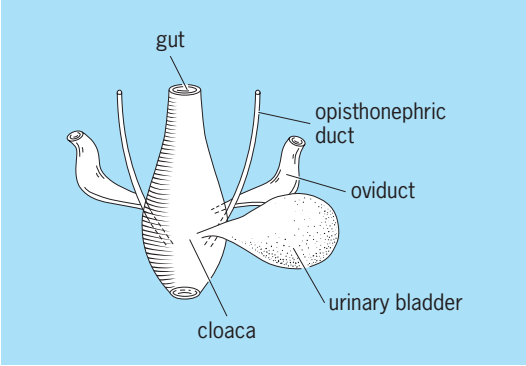


Fig. 14. Cloaca and urinary bladder of a female amphibian. (*After T. W. Torrey, Morphogenesis of the Vertebrates, 2d ed., John Wiley and Sons, 1968*)

Concomitantly, the openings of the excretory ducts are shifted and the urinary bladder is established. The pig embryo may serve to illustrate the usual events.

Almost as soon as the cloaca is established, an allantois arises from its floor. As in reptiles and birds, the allantoic sac, complete with blood vessels, expands greatly and fuses with the chorion. This association enables it to serve not only as a urinary bladder, but also to provide a major component of the placenta, so important in fetal economy. In the meantime the cloaca itself is modified. The cloaca initially ends blindly, its ventroposterior floor making contact with the embryonic skin to form the cloacal membrane (**Fig. 16***a*). A division of the cloaca into two parts, a dorsal rectum and ventral urogenital sinus, then follows. This division is effected by the urorectal fold, a crescentic fold which works backward from the angle where the allantois, excretory ducts, and gut meet until it meets the cloacal membrane (Fig. 16*b* and *c*). Not only is the cloaca itself subdivided, but the cloacal membrane is likewise divided into an anal and urethral membrane. Both these membranes eventually rupture so that the two parts of the cloaca open independently to the exterior.

The next step is the enlargement of that portion of the urogenital sinus receiving the neck of the allantois. The enlargement may involve a part of the allantoic stalk itself. This enlargement is the beginning of the definitive urinary bladder. As the bladder expands, the terminal ends of the Wolffian ducts which open to the urogenital sinus at this level are absorbed into the bladder wall. In consequence, the ureters, which stemmed initially from the Wolffian ducts, come to open directly to the growing bladder while the Wolffian ducts open somewhat behind into that part of the sinus which remains narrower and gives rise to the urethra (Fig. 16*c*). The ultimate fate of the Wolffian ducts and the final form of the urethra differ between the two sexes. The allantois per se is discarded at birth. Theodore W. Torrey

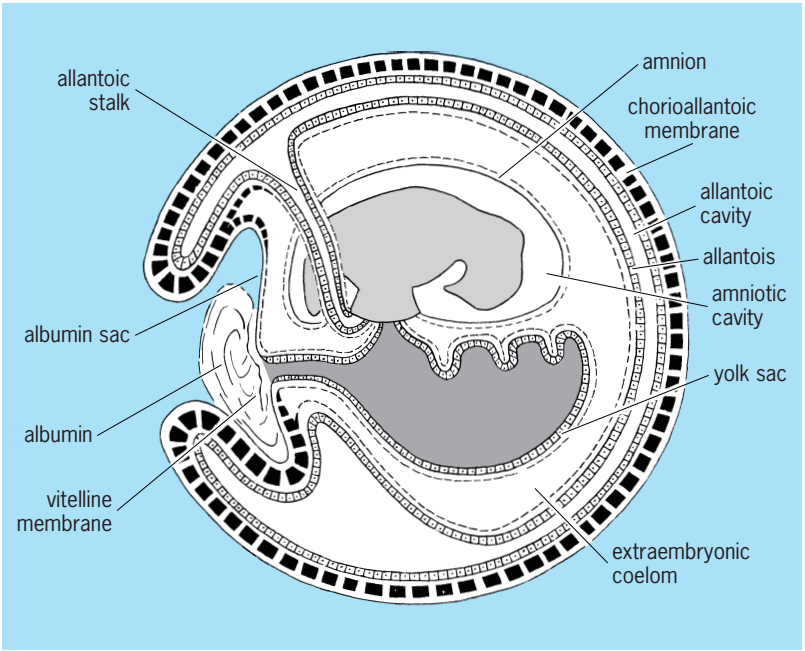


Fig. 15. Fully matured extraembryonic membranes of the chick. (*After T. W. Torrey, Morphogenesis of the Vertebrates, 2d ed., John Wiley and Sons, 1968*)

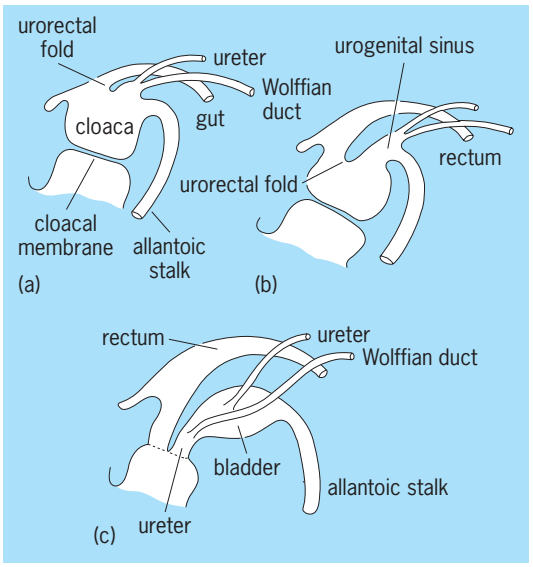


Fig. 16. Development of cloaca and urinary bladder. (*a*) Formation of cloacal membrane. (*b*) Urorectal fold formation. (*c*) Urethra formation.

Physiology

Urine is produced by individual renal nephron units which are fundamentally similar from fish to mammals (Fig. 4); however, the basic structural and functional pattern of these nephrons varies among representatives of the vertebrate classes in accordance with changing environmental demands. Kidneys serve the general function of maintaining the chemical and physical constancy of blood and other body fluids. The most striking modifications are associated particularly with the relative amounts of water made available to the animal. Alterations in degrees of glomerular development, in the structural complexity of renal tubules, and in the architectural disposition of the various nephrons in relation to one another within the kidneys may all represent adaptations made either to conserve or eliminate water. The urinary systems of fishes, amphibians, reptiles, birds, and mammals will be compared to see how they respond to demands imposed by various fresh-water, marine, amphibious, terrestrial, and desert environments that influence their development and differentiation.

General principles. Certain basic principles underlie excretory processes in all animals, from protozoa to humans. These are regulation of volume, electrolyte balance, movement of water across cell membranes, and the elimination of nitrogenous substances.

Regulation of volume. Paleontological evidence strongly suggests that early vertebrates evolved in fresh water, or at least that progenitors of modern fishes had a long history of dwelling in fresh water. Kidneys of living vertebrates bear the imprint of

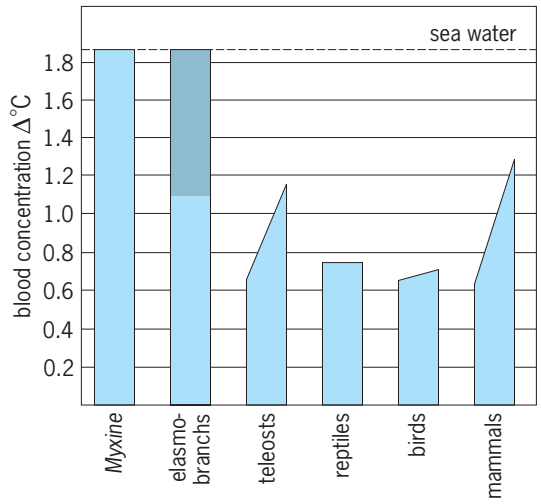


Fig. 17. Solute concentrations in blood of various marine vertebrates. Only *Myxine* and the elasmobranchs have blood isosmotic with seawater. Note the significant contributions of organic solutes, urea, and trimethylamine oxide (shaded) to the total osmotic concentration of elasmobranch blood. (*After A. P. McLockwood, Animal Body Fluids and Their Regulation, Harvard University Press, 1964*)

this early evolutionary history. Except for the primitive marine cyclostome *Myxine,* all modern vertebrates, whether marine, fresh-water, or terrestrial, have concentrations of salt in their blood only one-third or one-half that of seawater (**Fig. 17**). The early development of the glomerulus can be viewed as a device responding to the need for regulating the volume of body fluids. Hence, in a hypotonic fresh-water environment the osmotic influx of water through gills and other permeable body surfaces would be kept in balance by a simple autoregulatory system whereby a rising volume of blood results in increased hydrostatic pressure which in turn elevates the rate of glomerular filtration. Similar devices are found in fresh-water invertebrates where water may be pumped out either as the result of work done by the heart, contractile vacuoles, or cilia found in such specialized "kidneys" as flame bulbs, solenocytes, or nephridia that extract excess water from the body cavity rather than from the circulatory system. Hence, these structures which maintain a constant water content for the invertebrate animal by balancing osmotic influx with hydrostatic output have the same basic parameters as those in vertebrates that regulate the formation of lymph across the endothelial walls of capillaries.

None of the marine invertebrate phyla seem to have had a fresh-water ancestry. Members of the phylum Echinodermata, generally regarded as the closest relatives of the protovertebrate ancestors of fishes, have body fluids that are isosmotic with seawater, and they have no trace of any structure that resembles a kidney or other kind of excretory organ. *See* OSMOREGULATORY MECHANISMS.

Electrolyte balance. A system that regulates volume by producing an ultrafiltrate of blood plasma must conserve inorganic ions and other essential plasma constituents. The salt-conserving operation appears to be a primary function of the renal tubules which encapsulate the glomerulus. As the filtrate passes along their length toward the exterior, inorganic electrolytes are extracted from them through highly specific active cellular resorptive processes which restore plasma constituents to the circulatory system. The ability to form a urine hypotonic to blood is related to the length of the tubule. Animals, such as fresh-water fishes and amphibians, that produce a urine almost as dilute as distilled water have very long tubules with specially developed distal segments that are impermeable to water while they extract and restore to the blood almost every trace of salt from tubular urine before it enters the bladder. Proximal segments can also actively resorb salt, but their cells are freely permeable to water so that proximal tubular fluid is always isosmotic with respect to blood. Urine of marine fishes is always approximately isotonic with blood, and their renal tubules are short, with the specialized distal segment absent. Proximal tubule cells of all vertebrates also actively resorb glucose, amino acids, and other organic compounds from filtrate. Additionally they actively secrete foreign substances from blood into proximal tubular fluid.

Movement of water. Concentration gradients of water are attained across cells of renal tubules by water following the active movement of salt or other solute. Where water is free to follow the active resorption of sodium and covering anions, as in the proximal tubule, an isosmotic condition prevails. Where water is not free to follow salt, as in the distal segment in the absence of antidiuretic hormone, a hypotonic tubular fluid results. A special case is examined below of hypertonic urine formed when water is free to diffuse into a solution of high salt concentration surrounding mammalian collecting ducts when the cells are not correspondingly permeable to solute.

Nitrogenous end products. Of the major categories of organic foodstuffs, end products of carbohydrate and lipid metabolism are easily eliminated mainly in the form of carbon dioxide and water. Proteins, however, are more difficult to eliminate because the primary derivative of their metabolism, ammonia, is a relatively toxic compound. For animals living in an aquatic environment ammonia can be eliminated rapidly by simple diffusion through the gills. It is formed by relatively simple biochemical processes of deamination and transamination, which do not require the expenditure of free energy. However, when ammonia is not free to diffuse into an effectively limitless aquatic environment, its toxicity presents a problem, particularly to embryos of terrestrial forms that develop wholly within tightly encapsulated eggshells or cases. For these forms the detoxication of ammonia is an indispensable requirement for survival. During evolution of the vertebrates two energy-dependent biosynthetic pathways arose which incorporated potentially toxic ammonia into urea and uric acid molecules, respectively. Both of these compounds are relatively harmless, even in high concentrations, but the former needs a relatively large amount of water to ensure its elimination,

and uric acid requires a specific energy-demanding tubular secretory process to ensure its efficient excretion. *See* EXCRETION.

Methodology. The comparative physiological approach can be justified solely for information about the nature of the human's world, but it often contributes also to the solution of practical problems. For example, studies on the peculiar aglomerular kidney of the American anglerfish or goosefish (*Lophius*) led directly to the development of clinical renal clearance tests; micropuncture techniques that took advantage of the unusually large nephrons of the mudpuppy (*Necturus*) proved the hypothesis of glomerular filtration and identified specific sites along the renal tubule as loci of discrete secretory and resorptive processes; and observations on the structural and functional modifications in kidneys of the desert rat and other mammalian forms of arid regions showed how the countercurrent multiplier system operates in the renal medulla to concentrate urine. These examples show where an exploitation of experimental variables provided by nature led directly to new methods or information of general significance in renal physiology and pathology.

Rate of glomerular filtration. Following the disclosure that certain carbohydrates could not be excreted by aglomerular kidneys, a search was then undertaken to find some freely filterable nonmetabolizable carbohydrate which was also not resorbed by the renal tubules of glomerular kidneys. This led eventually to employment of the polysaccharide inulin in an application of the Fick principle to measure the rate of glomerular filtration. Knowing the total number of milligrams of inulin appearing in bladder urine per minute, one could determine the number of milliliters of filtrate needed to supply this amount by dividing the total number of milligrams excreted per minute by the number of milligrams of inulin in 1 ml of glomerular filtrate. This technique is applied to any intact animal or human patient by quantitatively collecting bladder urine, usually with the aid of a urethral catheter, and simultaneously sampling the inulin content of arterial blood. The latter is used in place of glomerular filtrate, which is virtually impossible to obtain, because the inulin contents of arterial blood and filtrate are identical. Only red blood cells and proteins are retained during the process of ultrafiltration. Hence the glomerular filtration rate (GFR) is determined by Eq. (1), where U_{in} is the amount

$$\text{GFR} = \frac{U_{in}V}{P_{in}} \tag{1}$$

of inulin in 1 ml of urine, V the number of milliliters of urine formed per minute, and P_{in} the amount of inulin in 1 ml of arterial plasma (filtrate).

Tubular resorption and secretion. Knowing the volume of filtrate formed per minute, one can measure the amount of any freely filterable plasma constituent being filtered as GFR $\times$ P_x, where P_x is the amount of plasma glucose, urea, amino acid, or other constituent in 1 ml of arterial blood (glomerular filtrate). The amount of this plasma constituent either resorbed or secreted by the tubules is then determined as the difference between the amounts filtered and that finally appearing in urine. For glucose, which is resorbed subsequent to filtration, this would be GFR $\times$ $P_G - U_GV$. For penicillinic acid, which is profusely secreted into urine by the renal tubule cells from postglomerular blood that supplied the tubules, it would be $U_{Pa}V -$ GFR $\times$ P_{Pa}.

For most organic compounds actively secreted or resorbed by the renal tubules subsequent to being filtered, there is a maximal limit imposed on the tubular transport rate (T_m). This maximal rate of transfer may be dependent upon the saturation of carrier capacity or upon a limitation in the amount of free energy available for moving these compounds across barriers against their own concentration gradients. Analogous compounds that are secreted or resorbed simultaneously may compete with one another for common carrier; raising the concentration of one will depress the transport rate of the other.

Renal plasma flow (RPF). The Fick principle is used to measure the amount of blood or plasma flowing through the kidneys per minute. For example, if the total milligrams of inulin excreted per minute ($U_{in}V$) is determined and divided by the number of milligrams removed from 1 ml of plasma as it traverses the kidney (arterial-venous difference or art. P_{in} − ven. P_{in}), then one knows the number of milliliters of plasma that was delivered to the kidneys per minute in order to account for the total number of milligrams of inulin excreted per minute.

However, use of inulin for this purpose calls for separate sampling of blood from the renal artery and the renal vein. Arterial blood from any part of the body can be used for concentrations of plasma constituents in renal arterial blood, but samples for venous blood must be from the renal vein, and this poses a serious problem in studies on intact animals or patients. Fortunately, certain foreign substances are secreted so profusely by the tubules that when they are infused into the circulation slowly, only low concentrations accumulate in arterial blood, and they are completely extracted while traversing the renal circulation. Inulin, excreted solely by glomerular filtration, is only 20% extracted in most mammals, but some compounds such as penicillin and *para*-aminohippuric acid (PAH) are almost completely removed by tubular secretion subsequent to glomerular filtration. So with low art. P_{PAH}, the value for ven. P_{PAH} may be assumed to be zero in the Fick equation, and the RPF is determined from Eq. (2). Thus the

$$\text{RPF} = \frac{U_{PAH}V}{\text{art. } P_{PAH}} \tag{2}$$

total number of milligrams of PAH extracted from the circulation by the kidneys per minute, divided by the number of milligrams of PAH extracted from each milliliter of plasma, gives the total number of milliliters of plasma that came to the kidneys per minute in order to account for the number of milligrams of PAH extracted (excreted) per minute. The small amount of nonextracted PAH in venous blood

probably represents that fraction of blood which supplies nontubular tissue. Hence, RPF measurements made by this method are sometimes referred to as "effective" RPF, to represent that which is selectively supplied to active tubular tissue; other extraglomerular blood goes to inert fat, connective tissue, and the renal capsule.

Filtration fraction. When both GFR and RPF are known, one can obtain another hemodynamic parameter. Alterations in the fraction of plasma filtered, FF = GFR/RPF, can give insight into whether vasomotor arteriolar changes occur mainly afferent to the glomeruli or as alterations in the diameter of efferent arterioles leading away from the glomeruli. If constriction occurred solely in the afferent glomerular arterioles, one would expect both GFR and RPF to fall correspondingly, and the FF would remain constant. If, however, resistance increased exclusively in the efferent arteriole, RPF would fall as before but the effective glomerular filtration pressure would be expected to rise, thereby increasing the fraction filtered.

In most mammals an average of about one-fifth of the plasma delivered to glomeruli is filtered. Lower vertebrates have much lower filtration fractions because of the separate renal portal (venous) circulation that bypasses the glomeruli and goes directly to the tubules. Hence, the high-pressure arterial blood supply that exclusively sustains glomerular filtration in all vertebrates constitutes a relatively small fraction of the total amount of blood delivered to the kidney in inframammalian forms.

Micropuncture. With the development of exquisitely delicate techniques that enable one to deal quantitatively with nanoliter (10^{-9} liter) fluid samples, physiologists turn to direct micrurgical techniques to study specific secretory and resorptive processes in assignable loci of renal tubule and collecting ducts. Minute samples of urine can be obtained either while urine is flowing freely within the tubular lumen, while tubules are being perfused with various solutions, or while flow is stopped by injections of oil at selective sites in "split-drop" experiments. Simultaneously, blood samples can be obtained from adjacent regions, and methods are available to measure electrical potential differences and flow of current across tubule cells or across separate cell membranes on the vascular and luminal sides of tubular epithelium.

Stop-flow technique. It is also possible to localize specific secretory and resorptive processes in intact animals by clamping the ureters to allow a stationary column of tubular urine to be held in contact with the tubular epithelium for several minutes. This permits cells in the various segments to operate on the urine in the column, thereby exaggerating processes that modify the composition of tubular urine in specific regions. The clamp is then released, and while urine spurts out under pressure, several dozen serial samples are collected and subjected to chemical analyses. The first samples come from the pelvis, collecting ducts, and distal segments of the tubule; later samples are from segments more proximal to the glomerulus; and finally filtrate and urine formed after release of the clamp are sampled. Some substance excreted solely by glomerular filtration is used as a marker, and its concentration reflects reabsorption of water by the tubule segments. Any urinary component whose concentration rises with reference to the marker is assumed to be secreted in that area, and when its relative concentration falls, it is assumed to be resorbed in the segment corresponding to that particular serially collected sample. **Figure 18** shows a stop-flow analysis in which the site of reabsorption of the amino acid glycine AN is identified as being in the same proximal region in which PAH is secreted. Ammonia (NH_4^+) is secreted distally. In this case creatinine was used as a marker, and the urine/plasma (U/P) concentration ratios of PAH and NH_4^+ were factored by the corresponding U/P ratio of creatinine to correct for reabsorption of water which occurred at varying rates along the nephron during the stoppage of flow.

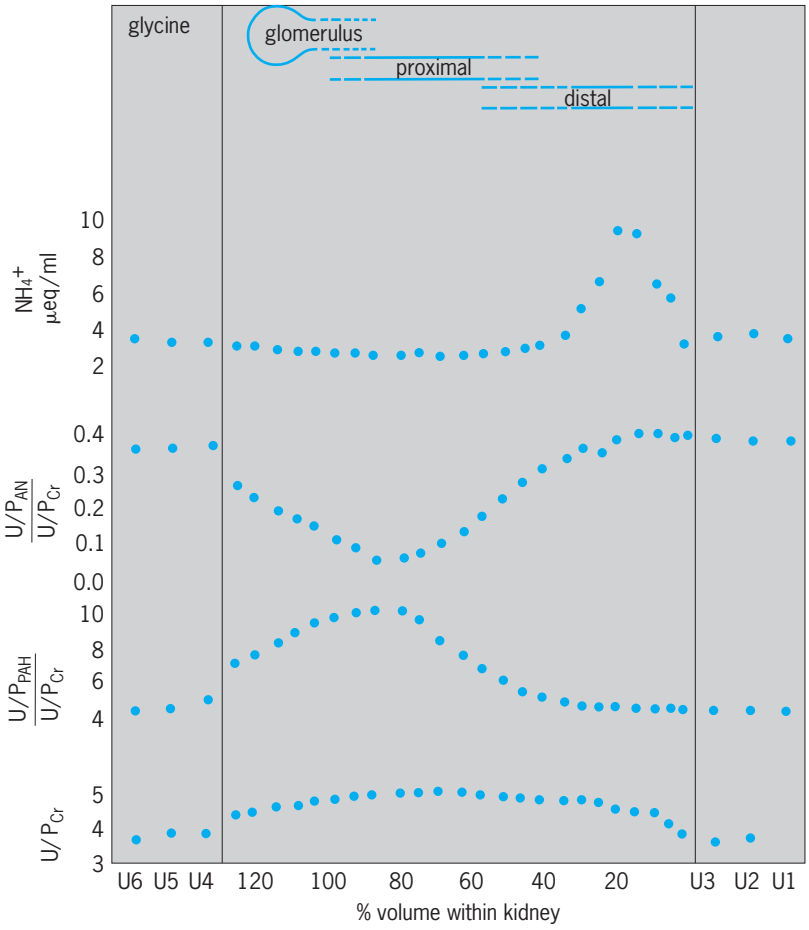


Fig. 18. Stop-flow analysis of the sites of ammonia secretion, aminonitrogen resorption, and PAH secretion. (*After R. F. Pitts, Physiology of the Kidney and Body Fluids, Year Book Medical Publishers, 1963*)

Culture methods. Isolated renal tubules and thin slices of kidney in suitably oxygenated, balanced salt solutions containing substances capable of being secreted continue to carry on their cellular transport activities for several hours. In the simplest preparations a colored organic compound such as phenol

red can be directly observed under the microscope as it is progressively concentrated to a very high degree within the lumen of renal tubules. Such studies have thrown considerable light on carrier-mediated transport competition and the nature of energy dependence. More complicated techniques involve use of chemical analyses and techniques for microperfusing isolated tubules with synthetic solutions of known composition. Fish kidneys, especially of the marine flounder, are particularly useful for tissue culture because they contain little cementing substance and the individual tubules can be easily teased apart. *See* TISSUE CULTURE.

Patterns of excretion. Nephrons of primitive protovertebrates probably were supplied exclusively with venous blood, and the open-ended tubules had nephrostomes that drained fluid directly from the body cavity. With the assumption of a fresh-water habitus and the need for osmoregulation, it is thought that fishes acquired a glomerulus that at first was perhaps only loosely associated with the coelomostome, as in the embryonic pronephros of present-day vertebrates. Eventually a dual blood supply became possible when the glomerulus supplied with arterial blood became tightly encapsulated within the end of the tubule, but the open coelomostome persisted in some lower vertebrates. With the evolution of a homeostatically regulated internal environment and the guarantee of constant high-pressure filtration at the glomerulus, the renal portal system disappeared in mammals leaving the renal tubule solely with an arterial postglomerular bloody supply (**Fig. 19**).

The functional significance of the venous renal portal supply in a cold-blooded form such as the frog, for example, probably derives from its ability to maintain tubular secretory activity under environmental conditions that would tend to reduce arterial blood pressure to the point where glomerular filtration ceases. Glomerular activity generally is more labile in cold-blooded vertebrates than in mammals. Rates of filtration rapidly fall in fishes and amphibians when they are stressed either by evaporation or osmoregulatory water deprivation, or by cold. Lower vertebrates also rely more on tubular secretory processes for sustenance when glomerular filtration is diminished or ceases entirely and when the low-pressure venous supply of the renal portal system may be the sole source of blood for the kidney.

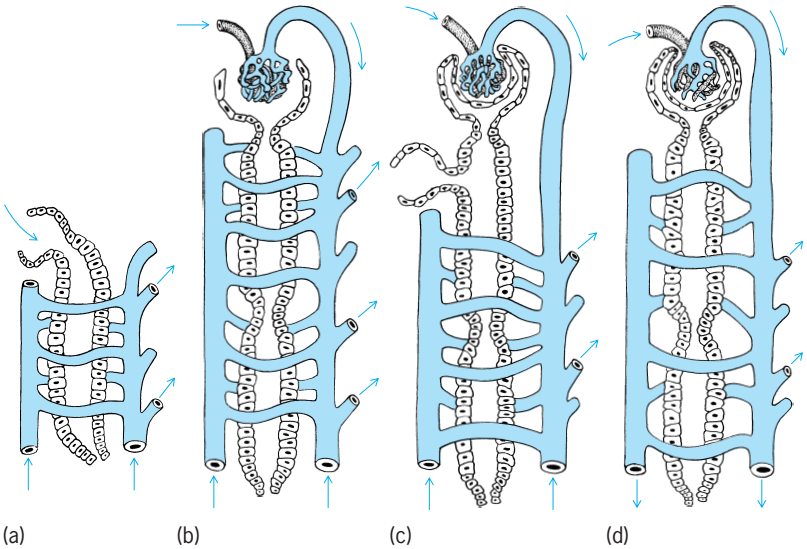


Fig. 19. Probable stages in the evolution of the vertebrate nephron; arrows indicate direction of flow. (*a*) Hypothetical protovertebrate tubule drained the body cavity through an open coelomostome with only a venous blood supply. (*b*) In earliest vertebrates the coelomostome was brought into loose association with knots of arterial blood vessels capable of filtering fluid under hydrostatic pressure. (*c*) Later evolution sealed Bowman's capsule in the glomerulus. (*d*) In mammals the renal portal venous supply and coelomostomes disappeared. (*After H. W. Smith, From Fish to Philosopher, Little, Brown, 1959*)

Teleost fishes. The urinary system of fishes plays only a secondary role in the excretion of nitrogenous waste products. As much as 90% of the nitrogen that comes from the deamination of amino acids is usually excreted at the gills in the form of ammonia. Fish in fresh water form large quantities of glomerular filtrate, and this is desalinated by their characteristically long renal tubules which eventually produce copious volumes of essentially salt-free urine. Fresh-water fishes excrete urine equivalent to 30% of their body weight per day, and the dilute urine may have freezing point depressions as low as Δ 0.07°F (Δ 0.04°C). Anadromous forms such as the salmon and steelhead trout show markedly reduced urine flow and glomerular filtration rates when in their seawater habitat.

Marine teleosts need to conserve free water in maintaining solute concentrations in blood that are only about half that of seawater (average is approximately Δ 1.62°F (Δ 0.9°C); seawater freezes at 28.54°F (−1.86°C). Production of urine is scanty; glomeruli are reduced in size and totally absent in some 20 species of marine teleosts. After intake of seawater and following intestinal absorption, blood is desalinated to maintain water balance by the gills, which remove the monovalent sodium and chloride ions, and the kidneys, which extract divalent magnesium, calcium, and sulfate ions. These ion transfers are carried out by specific energy-demanding active transport processes, which operate in opposite directions in similar branchial and renal systems of fresh-water fishes. Distal tubules are absent in marine fishes, and the urine of both the glomerular and aglomerular forms is roughly isosmotic with that of blood.

Cyclostomes. Primitive hagfishes such as *Myxine* resemble marine invertebrates and are unique among vertebrates in that their blood is isosmotic with seawater, due almost entirely to the accumulation of inorganic ions (Fig. 17). Hence no osmoregulatory mechanisms are required. Fresh-water and brackish lampreys such as *Lampetra*, on the other hand, produce copious flows of dilute urine, and their renal function resembles that of fresh-water and diadromous teleosts.

Cartilaginous fishes. Sharks, skates, rays, and chimeroids living in the sea prevent water loss to the marine environment in a way that is totally different from that of marine teleosts. They have relatively

large glomeruli and they do not drink seawater; instead the renal tubules have specialized processes capable of extracting urea and trimethylamine oxide from glomerular filtrate, and the active resorption of these organic compounds raises the osmotic concentration of their blood and body fluids to equality with seawater (Fig. 17).

Lungfishes. Urinary systems of fresh-water dipnoans function similarly to those of fresh-water teleosts. However, during estivation all urinary functions stop, and potentially toxic ammonia is then converted by the liver to urea which may accumulate in body fluids to extremely high levels without causing damage. When floods occur after periods of drought sometimes lasting several years, water is rapidly taken up osmotically, and the resulting increase in blood volume and pressure restores glomerular filtration; then a copious production of urine rapidly removes urea and other waste products accumulated during estivation. It is interesting that of the three surviving lungfishes, the Australian lungfish (*Neoceratodus forsteri*), which cannot estivate or survive out of water, is the only one incapable of synthesizing urea via the so-called ornithine cycle that accounts for the production of urea in the African and South American lungfishes, both of which can estivate.

Amphibians. The purely aquatic urodeles, such as *Necturus*, and the tadpole stages of anuran frogs and toads have patterns of urinary excretion similar to that of fresh-water teleosts. However, during metamorphosis when anurans change from gill to lung breathing, nitrogen is excreted principally in the form of urea rather than ammonia. Urea is actively secreted by the tubules of the more aquatic frogs, but it does not appear in terrestrial toads or in the unusual crab-eating frog (*Rana cancrivora*) of Southeast Asia that lives in marine mangrove swamps. Such retention of urea could serve the same osmoregulatory function that it does in marine cartilaginous fishes, that is, to help toads in an arid environment to absorb water quickly through the skin during brief showers, and to provide a more favorable free-water gradient for the marine toad in its hyperosmotic environment. Renal accommodations to semiterrestrial life in amphibians, which involve glomerular recruitment, facultative tubular reabsorption of water, and alterations in the synthesis and secretion of urea are undoubtedly implicated also in such amphibious fishes as the tree-climbing perch (*Anabas*), the mudskippers, many gobies, and such reptiles as the aquatic turtles, alligators, and many water snakes which may alternately live in and out of an aqueous environment.

Terrestrial reptiles and birds. These egg-laying forms have adapted to "uricotelism," which effectively provides for detoxification of ammonia and also for extremely efficient conservation of water. In addition to having relatively low rates of glomerular filtration, uric acid can be actively secreted by the tubules to form a urine practically saturated with urate. Actual precipitation of solid urate occurs in the allantoic bladder of embryos and in the cloaca of adults, and masses of white crystals are found as residues in hatched eggshells and feces, respectively.

Mammals. The placental form of embryonic development in mammals guaranteed an aquatic environment for the fetus which permitted retention of the "ureotelic" habitus. The renal portal system disappears in mammals, whose tubules are supplied solely with blood that has previously traversed the arterial glomerular capillaries (Fig. 19). The unique functional feature of the mammalian kidney is its ability to concentrate urine. Human urine can have four times the osmotic concentration of plasma (1200 milliosmoles/liter), and some desert rats that survive on a diet of seeds without drinking any water have urine/plasma concentration ratios as high as 17 (6.34 Δ osmoles/liter, more than six times that of seawater). More aquatic forms such as the beaver have correspondingly poor concentrating ability.

The concentration operation depends on the existence of a decreasing gradient of solute concentration that extends from the tips of the papillae in the inner medulla of the kidney outward toward the cortex. The high concentration of medullary solute is achieved by a double hairpin countercurrent multiplier system which is powered by the active removal of salt from urine while it traverses the ascending limb of Henle's loop (**Fig. 20**). The salt is redelivered to the tip of the medulla after it has diffused back into the descending limb of Henle's loop. In this way a hypertonic condition is established in fluid surrounding the terminations of the collecting ducts. Urine is concentrated by an entirely passive process as water

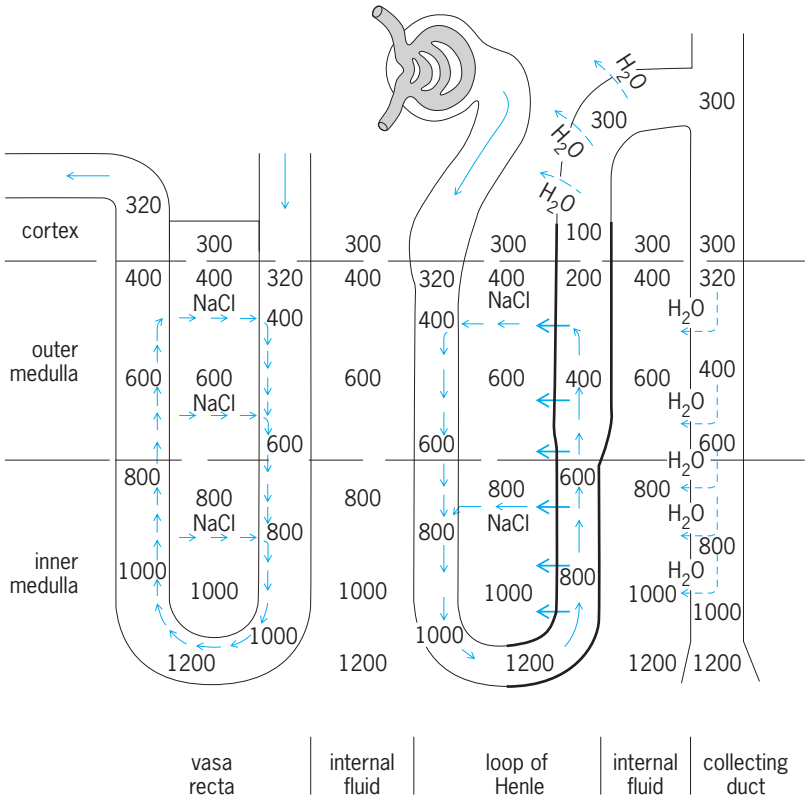


Fig. 20. The tubular countercurrent mechanism for concentrating the urine, and the exchange system in the straight hairpin blood vessels of the medulla, which minimizes loss of solute from medullary interstitial fluid. Arrows indicate direction of movement of solute. (*After A. C. Guyton, Textbook of Medical Physiology, Saunders, 1966*)

leaves the lumen of collecting ducts to come into equilibrium with the hypertonic fluid surrounding its terminations.

Urine more dilute than plasma is formed when hypotonic urine delivered by the ascending limb to the cortex traverses the lengths of distal tubules and collecting ducts and it is not free to come into equilibrium with the progressively more salty interstitial fluids of the cortex, outer medulla, and inner medulla, respectively.

Another countercurrent exchange system operating in the straight blood vessels (vasa recta) that parallel the loops of Henle prevents washout of the hypertonic interstitial fluid of the inner medulla where the blood is in osmotic equilibrium with the surrounding interstitial fluid. Because of the hairpin arrangement of the vasa recta, salt and other solutes are trapped and kept from being carried away from the medulla (left side of Fig. 20). As solute diffusing from the ascending capillary limb comes into equilibrium with more dilute interstitial fluid nearer the cortex, it enters the descending vessel to be redelivered to the inner medulla. This entirely passive process involving exchange of solute is the same as that operating in countercurrent heat exchangers.

Removal of most of the salt from glomerular filtrate takes place in the proximal convoluted tubules of the renal cortex. Sodium is actively resorbed from tubular fluid, and chloride then passively follows down an electrical gradient. Removal of solute effects the simultaneous isosmotic entry of water into the cortical interstitial fluid. This solution, containing about 300 milliosmoles of solute, is then rapidly restored to the general circulation by blood perfusing the cortical capillaries. In humans, about 80% of the glomerular filtrate is withdrawn by the time tubular fluid enters the descending limbs of Henle's loops. During water diuresis the distal tubules and collecting ducts are relatively impermeable to water, and the hypotonic urine then leaving the loops of Henle is made even more dilute by continued active extrusion of ions, finally resulting in the elimination of large volumes of urine containing little solute. In contrast, during conditions demanding conservation of water antidiuretic hormone (ADH) is released by the posterior pituitary gland, and the distal tubules and collecting ducts are then rendered freely permeable to water. Urine in the presence of ADH is isotonic by the time it reaches the middle of the distal segment, and its volume has been drastically reduced by the continued active extrusion of sodium and the associated passive diffusion of water. The small amount of isosmotic urine entering the collecting duct is progressively concentrated as it gives up water to the hypertonic medullary interstitial fluid and finally achieves a maximal concentration of solute equivalent to that of fluid surrounding the collecting ducts at the papillary tips.

Regulation mechanisms. Many years ago it was shown that injection of crude extracts of pituitary glands into frogs and toads caused rapid weight gain by accumulation of water. Now a whole family of octapeptides is known to be secreted by neurons of the hypothalamus of the brain and then transported by axons to the pituitary gland where they are stored in the neurohypophyseal lobe. Two of these, arginine vasotocin and vasopressin, oppose diuresis in terrestrial and semiterrestrial vertebrates, and they are frequently referred to as ADH. ADH increases water uptake by opening pores in the skin and urinary bladders of anurans and diminishes urinary output both by increasing the extraction of water from distal tubular urine and by reducing the rate of glomerular filtration. Secretion of ADH is regulated by osmoreceptors in the hypothalamus. Among teleosts secretions of a neurosecretory system in the posterior region of the spinal cord are passed along into a neurohemal organ, the urohypophysis, via a structural and functional arrangement similar to the hypothalamic-neurohypophyseal system of the terrestrial forms.

Renal regulation of electrolytes is accomplished by steroid hormones produced by the adrenal cortex or by homologous tissues in the lower vertebrates. In the absence of cortical steroids kidneys cannot adequately retain sodium and chloride; serum potassium levels are elevated; and there is a marked reduction in blood volume with death inevitably ensuing. Aldosterone is the most potent cortical steroid in mammals. Its release is regulated by changes in the electrolyte level of blood flowing through the brain and kidney. It also promotes sodium retention and potassium loss in sweat and gastrointestinal glands. *See* STEROID.

In the frog, however, aldosterone induces the renal tubule to retain potassium and lose sodium. Its effect on the kidney is just the opposite of that in mammals; nevertheless, the frog recovers sodium even more efficiently from bladder urine, and the net result in conserving salt is the same.

Renin is a proteolytic enzyme probably produced by specialized secretory cells in the afferent glomerular arteriole. It reacts with a plasma component to produce a nonprotein substance called angiotensin II which is capable of raising arterial blood pressure. There are several lines of evidence that suggest that the kidney's renin-angiotensin system is a prime regulator of the secretion of aldosterone by the adrenal cortex. Apparently some kind of feedback regulation mutually controls these two important secretions that regulate blood pressure, renin by affecting peripheral resistance in the arterioles, and angiotensin II by controlling retention of salt, hence blood volume and pressure. A negative feedback mechanism could be triggered by hemorrhage or some other factor, causing a temporary decrease in renal arterial blood pressure and renal blood flow. The resultant hypotension could activate the renin-angiotensin system causing the release of aldosterone, which in turn would induce the renal tubules to retain sodium and to increase blood volume, and consequently to raise blood pressure. Such a negative feedback path would then be the afferent signal needed to inactivate the renin-angiotensin system. This is speculative, but the system is of interest to both clinicians and comparative physiologists.

There is also a feedback regulatory system within the kidney itself that apparently coordinates tubular and glomerular function. The juxtaglomerular apparatus which is found in all classes of vertebrates consists of specialized tubular epithelium at the end of Henle's loop (macula densa) and of renin-containing cells in the afferent arterioles. These two separate types of specialized cells come into intimate contact with one another, and the two apposing parts always belong to the same nephron. There is considerable evidence to support the hypothesis that an increased concentration of salt in tubular urine at the macular densa site activates a sodium-sensitive feedback system which operates through the juxtaglomerular apparatus to reduce the rate of glomerular filtration and thereby adjust the tubule's load of filtered salt to the sodium reabsorptive capacity of this particular nephron. *See* ENDOCRINE MECHANISMS; URINE.

Roy P. Forster

Bibliography. R. P. Forster, Comparative vertebrate physiology and renal concepts, in *Handbook of Physiology*, sect. 8: *Renal Physiology*, pp. 161–184, American Physiological Society, 1973; D. J. Griffiths, *Urodynamics*: *The Mechanics and Hydrodynamics of the Urinary System*, 1980; A. C. Guyton, *Textbook of Medical Physiology*, 10th ed., 2000; R. Hill and G. Wyse, *Animal Physiology,* 2d ed., 1989; W. S. Hoar, *General and Comparative Physiology*, 3d ed., 1983; C. L. Prosser, *Comparative Animal Physiology*, 4th ed., 1991; T. W. Torrey, *Morphogenesis of the Vertebrates*, 4th ed., 1979.

Urinary tract disorders

Diseases affecting the outflow tract of the urinary system from the renal pelvis to the distal portion of the urethra. Disorders of the kidney itself are considered in a separate article. *See* KIDNEY DISORDERS.

Structures and Associated Disorders

Urinary tract disorders can occur in the renal pelvis, ureters, bladder, or urethra.

Renal pelvis. The renal pelvis, the structure that collects and carries urine from the kidney to the ureter, is subject to a number of disorders.

Anomalies. Anomalies of the renal pelvis are common. The pelvis may be subdivided to form a duplicated system, which may join a single ureter or empty into separate ureters. By itself such an anomaly is harmless, but if there is associated obstruction or vesicoureteral reflux (abnormal flow of urine from the bladder back into the kidney), there may be symptoms of pain, associated infection, and possibly destruction of renal tissue.

Calculi. Calculi typically arise in the renal papillae usually below the epithelial layer. The cause is usually due to a metabolic defect, which results in excessive excretion of oxalate, urate, calcium, and/or cystine. The most common stones are calcium-containing, such as calcium oxalate and calcium phosphate. Precipitating factors include dehydration, dietary excess of certain foods, and infection. Also, in the presence of a bacterial infection or foreign material, the salts, which are often in a supersaturated condition in the urine, may precipitate out. Thereafter, the calculi enlarge by accretion of the same or other salts.

After reaching a certain size, calculi break through the epithelial layer and migrate into the renal pelvis and float free. They may pass down the ureter. Some stones, particularly magnesium ammonium phosphate calculi (struvite stones), are associated with bacterial infection and may become very large and take the shape of the renal pelvis and infundibulae. These are, therefore, called staghorn calculi. In addition to causing discomfort, they can also cause obstruction and infection of the renal parenchyma, which can lead to loss of kidney function. Patients with staghorn calculi must be treated to remove the stones.

Other obstructions. Fibrous bands and aberrant vessels associated with the kidneys are usually blamed for obstruction when no other intrinsic cause can be found for ureteropelvic obstruction. In some cases, there may be intermittent obstruction from "sagging" of the kidney in the upright posture. Tumors, either intrinsic or extrinsic, and periureteral fibrosis (fibrosis around ureter) are also capable of blocking the flow of urine from the renal pelvis.

In complete acute obstruction from any cause, the backpressure in the collecting system hinders urine secretion from the kidney. This process is reversible if the obstruction is relieved. In chronic partial obstruction, the most common type, the backup of urine in the renal collecting system leads to dilation of the pelvis, known as hydronephrosis. Hydronephrosis is classified as mild, moderate, or severe and reflects the degree of dilation of the renal pelvis due to the obstruction.

Obstruction causes urinary stasis and this contributes to developing infection (pyelonephritis; if pus develops the condition is known as pyonephrosis). The resulting distension and (often) infection may destroy some or all of the renal parenchyma, with deleterious effects on renal function. The kidney may ultimately become an abnormally enlarged sac with no useful function.

Infection. Acute infection of the kidney, known as pyelonephritis, produces fever and flank pain. If there are no underlying causes (such as stone, obstruction, or tumor), the disease can be treated with antibiotics and has no serious aftereffects.

Tumors. Tumors of the pelvis are rare; when they do occur they are usually transitional-cell carcinomas with varying degrees of differentiation. They can cause obstruction and hematuria. They are often associated with similar tumors elsewhere in the urinary system. It is not known whether this represents superficial implantation of tumor or new tumor growth due to persistence of inciting factors.

Ureters. These long thin tubes that carry urine from the renal pelvis to the bladder are subject to the same diseases as the renal pelvis.

Anomalies. Duplication of the ureters can occur at any point, from a dual pelvis and fusion to form one ureter, to complete duplication and separate

implantation into the bladder. The anomaly is generally harmless unless obstruction or reflux is present, which can contribute to developing infections. Sometimes an anomalous ureter inserts into the vagina and causes dribbling of urine, for which surgical correction is required due to incontinence.

Calculi. One of the most common causes of excruciating abdominal and flank pain, often radiating to the groin or testicle, is the passage of small calculi. They originate from the renal pelvis and travel toward the bladder. If the stone becomes lodged along the course from the kidney to the urethra, intervention is required to remove the stone. If the stone causes obstruction, the ureter and renal pelvis proximal to the stone becomes dilated (hydroureteronephrosis). Again, obstruction of the urinary system can lead to renal damage and loss of renal function.

Other obstructions. Occasionally, the ureter is obstructed by external pressure from retroperitoneal fibrosis (development of a fibrous mass in the back of the abdomen behind the abdominal lining), tumor, or aortic aneurysm. In addition, any chronic obstruction of the lower urinary tract (bladder or urethra) will lead to progressive dilation of the ureters and renal pelvis. The effect is the same no matter what the cause; obstruction can lead to permanent renal damage if not relieved.

Vesicoureteral reflux. If the normal valvelike mechanism that prevents urine in the bladder from going back toward the kidney is defective, vesicoureteral reflux results. Reflux can occur during bladder filling or emptying. When associated with an infection, vesicoureteral reflux can lead to renal injury.

Tumors. These are usually transitional-cell carcinomas and do not differ appreciably from those seen in the renal pelvis and bladder.

Bladder. The urinary bladder is subject to anomalies, obstruction, inflammation, calculi, fistulae, and tumors. As the receptacle for urine from the kidneys, it is variable in shape and form depending upon the amount of urine it contains.

Anomalies. Bladder exstrophy is an uncommon congenital anomaly due to failure of both the lower abdominal wall and anterior bladder wall to close during development. The bladder appears directly through the abdominal wall defect. Occasionally a fetal structure, such as the urachus, fails to close by the time of birth and drains urine through the umbilicus. Abnormal congenital valves of the posterior urethra, known as posterior urethral valve, interfere with bladder emptying. This condition occurs only in males and is significant because it can lead to chronic renal obstruction and subsequent renal failure.

Calculi. Bladder calculi usually develop as a result of infection or obstruction. They can be of varying size and are often composed of phosphates and urates. After removal or spontaneous passage of stones, new stones tend to form unless the predisposing factors are corrected.

Obstruction. Interference with complete emptying of the bladder may be secondary to various factors. In males it is often due to prostatic enlargement, and in females it is typically due to a cystocele, which is a bulging of the bladder into the vaginal vault. Other causes of obstruction include calculi; tumors; constriction secondary to surgical instrumentation or gonococcal infection; or congenital valves. A neurogenic bladder (which occurs when the normal balance of bladder wall contraction and sphincter relaxation is altered and the bladder cannot completely empty) usually follows spinal cord injury or disease and can lead to inadequate bladder emptying.

Obstruction from any cause leads to changes in the bladder wall. The bladder wall can weaken in areas and bulge outward, causing the formation of a diverticulum. Other effects of obstruction include increased susceptibility to infection, stone formation, and backpressure on the kidneys and ureters.

Inflammation. Cystitis or inflammation of the urinary bladder is a common affliction and is characterized by dysuria (painful urination), frequency, and urgency. The organisms responsible are primarily bacterial, including *Escherichia coli*, pseudomonas, and other gram-negative organisms, although yeast and fungi can cause urinary tract infections. Tubercle bacilli are rarely involved. Bacteria can gain access to the bladder from the blood or, more commonly, ascend from the urethra. The lesions include granular or even cystic thickening of the epithelium and redness and swelling of the wall. A peculiar plaquelike lesion called malakoplakia, which is usually associated with the bladder but can be found elsewhere in the body, is known to be a reaction to *E. coli*.

Fistulas. Abnormal connections (fistulas) between the bladder and the female urethra or vagina are not unusual. These can be due to pelvic or transvaginal surgery, childbirth, or radiation. Enterovesical fistulas (abnormal connection between the bowel and bladder) can also be secondary to pelvic surgery, radiation therapy, or inflammatory bowel diseases. Fistulas can result in incontinence and chronic urinary tract infection.

Tumors. The bladder is frequently the site of tumors, which are almost always of transitional-cell origin. Although benign lesions can exist in the bladder, bladder cancer needs to be ruled out. Bladder cancer can remain superficial or it can penetrate the bladder muscle, invade neighboring structures, obstruct the ureters or urethra, and metastasize to distant organs. Malignant tumors are more frequent in cigarette smokers and in people who work near aniline dyes. These observations indicate that carcinogens excreted in the urine play a role in initiating bladder cancer.

Female urethra. This structure is not often diseased. In older women a painful raspberrylike growth of inflammatory tissue may appear at the external urethral opening and is known as a caruncle. The urethra can also be a site of transitional-cell and squamous cell malignancy.

Roy N. Barnett; Katherine S. Rhee; Martin I. Resnick

Male urethra. The male urethra is approximately 20 cm in length and is responsible for draining urine from the bladder and acting as a conduit for the delivery of semen. The first 3 cm is surrounded by the prostate gland and is termed the prostatic urethra.

The prostate gland may develop cancer, and options are available to treat this condition based on stage of the tumor or extent of disease at time of detection.

The prostatic urethra is prone to obstruction. The most common cause is a noncancerous growth of the prostate called benign prostatic hyperplasia. This growth is a natural process associated with aging and can occlude the prostatic urethra, causing progressive difficulty in urinating. Individuals may complain of straining with urination, weak stream, incomplete bladder emptying, nocturia (excessive nighttime urination), frequency, and/or double voiding. Similar symptoms may be secondary to urethral strictures which can be secondary to trauma, surgical instrumentation, or gonococcal infections. Symptoms of benign prostatic hyperplasia can be treated with medications to shrink the prostate or relax the bladder neck. If symptoms are severe or there is an inability to void, a portion of the prostate needs to be removed surgically. Less-invasive procedures such as microwave therapy, lasers, and transurethral needle ablation can be used as well. If the prostate is too large for an endoscopic procedure, it can be removed by open surgery.

Signs and Symptoms

Signs and symptoms of urinary tract disorders may be generalized or localized. General disturbances include fever and pain. Malaise can accompany fever, but it may also be due to the body's response to the toxins normally eliminated by the kidney if there is loss of kidney function. Local signs and symptoms depend upon the location of the disease. When the renal pelvis or ureters are obstructed or infected, back or abdominal pain overlying the affected organ often arises. The pain can be severe and radiate down to the groin or testicle. Other urinary tract disorders, such as bladder cancer, can present with asymptomatic hematuria (presence of blood in the urine).

Diagnosis

Techniques used to diagnose urinary tract disorders include clinical laboratory analysis of blood and urine, radiological procedures, and endoscopy.

The most common clinical laboratory procedure is urinalysis, which involves microscopic examination of the urine to identify red and white blood cells not normally present. If infection is suspected, cultures for bacteria can be done. Serum studies are useful in assessing renal function and identifying metabolic disorders. *See* URINALYSIS.

Intravenous contrast is used to help delineate and examine the urinary tract with x-rays. Stones can be diagnosed with a plain abdominal x-ray or computed tomography without contrast. In certain situations, radioisotopic dyes can be injected to determine the function of each kidney. The urinary tract can also be examined with ultrasound. *See* MEDICAL IMAGING.

Endoscopy requires the use of special instruments to directly visualize the urethra, bladder, ureters, and renal pelvis. Radiopaque dyes can also be injected into the urinary tract by this approach. When abnormal tissue is visualized, biopsies can be obtained and sent for pathologic examination.

Treatment

Treatment depends upon the type of urinary tract disorder that is present. For infections, a urine culture and sensitivity assessment should be obtained and a course of oral or intravenous antibiotics administered and adjusted appropriately according to the organism identified.

Obstructive lesions, such as benign prostatic hyperplasia, can be treated surgically or with medication to relieve lower urinary tract symptoms. Surgical intervention includes endoscopic or open procedures to remove excess prostate tissue. Treatment with medication involves the use of α-blockers and 5-α reductase inhibitors.

Stones are treated in a variety of ways including administration of medication and/or surgical management. Patients can choose to wait for the stones to pass spontaneously or can elect for surgical intervention, including electrohydraulic shock wave lithotripsy (ESWL®), ureteroscopy with lithotripsy or basket retrieval, percutaneous surgery, or open surgery.

The treatment of tumors, for example in the bladder, requires resection (surgical excision) to determine the pathological diagnosis and also the depth of invasion. If a bladder tumor is superficial, tumor resection can be followed with intravesical immunotherapy or chemotherapy to prevent recurrent tumors. More invasive disease should be treated with radical cystectomy, which involves removal of part of the urinary bladder, as well as the prostate (in men) and the uterus, fallopian tubes, ovaries, anterior vaginal wall and the urethra (in women), along with bilateral pelvic lymph node dissection. Combined radiation and chemotherapy is also an option in patients with invasive disease. Tumors of the upper tract may require nephroureterectomy (surgical removal of the kidney, ureter, and top part of the bladder) or distal ureterectomy (removal of the distal portion of the ureter) with ureteral implant, depending upon the tumor location and grade. *See* CANCER (MEDICINE); REPRODUCTIVE DISORDERS; TUMOR; URINARY SYSTEM.

Katherine S. Rhee; Martin I. Resnick

Bibliography. M. I. Resnick et al., *Critical Decisions in Urology*, 3d ed., 2004; P. C. Walsh et al., *Campbell's Urology*, 8th ed., 2002.

Urine

An aqueous solution of organic and inorganic substances, mostly waste products of metabolism, which collects in the kidneys and is removed from the body primarily through the bladder and urinary tract. The kidneys maintain the internal milieu of the body by excreting these waste products and adjusting the loss of water and electrolytes to keep the body fluids relatively constant in amount and composition.

About 1.5 quarts (1500 ml) of blood courses through the kidneys each minute. As it passes through the glomeruli, the pressure of the blood, opposed by the osmotic pull of the plasma proteins, tends to drive water and salts through the wall of the glomerular tuft into the glomerular space at a rate of 7.2 $in.^3$ (120 ml) per minute. This dilute glomerular filtrate courses through the kidney tubules where it is concentrated by active reabsorption of water, glucose, and essential salts. Approximately 119 parts of 120 of the total volume is reabsorbed, leaving only 1 part to be excreted as urine which, in the adult, under normal circumstances, amounts to approximately 1.2–1.5 quarts (1200–1500 ml) per day. *See* OSMOREGULATORY MECHANISMS.

The urine normally is clear and has a specific gravity of 1.017–1.020, depending upon the amount of fluid ingested, perspiration, and diet. The increase in specific gravity above that of water is due to the presence of dissolved solids, about 60% of which are organic substances such as urea, uric acid, creatinine, and ammonia; and 40% of which are inorganic substances such as sodium, chloride, calcium, potassium, phosphates, and sulfates. Its reaction is usually acid (pH 6) but this too varies with the diet. It usually has a faint yellow color due to a urochrome pigment, but the color varies depending upon the degree of concentration, and the ingestion of certain foods (for example, rhubarb) or cathartics. It usually has a characteristic aromatic odor, the cause of which is not known. *See* URIC ACID.

The chemical and physical composition of the urine frequently reflects pathological changes not only in the genitourinary tract but disease processes in other parts of the body. *See* CLINICAL PATHOLOGY; KIDNEY; URINALYSIS; URINARY SYSTEM.

Reuben Straus

Urodela

One of the orders of the class Amphibia, also known as the Caudata. The members of this order are the tailed amphibians, or salamanders, and are distinguished superficially from the frogs and toads (order Anura) by the possession of a tail, and from the caecilians (order Apoda) by the possession of limbs.

Morphology. In appearance, salamanders resemble lizards in that members of both groups normally have a relatively elongate body, four limbs, and a tail. The similarity is, however, only superficial. The most obvious external difference, the moist glandular skin of the amphibian and dry scaly skin of the reptile, is underlain by the numerous characters that distinguish reptiles from amphibians.

The vast majority of salamanders are well under a foot in length. The largest is the giant salamander of Japan and China, which may attain a length of 5 ft (1.5 m). A close relative, the hellbender, which lives in streams in the eastern United States, grows to over 2 ft (0.6 m) in length.

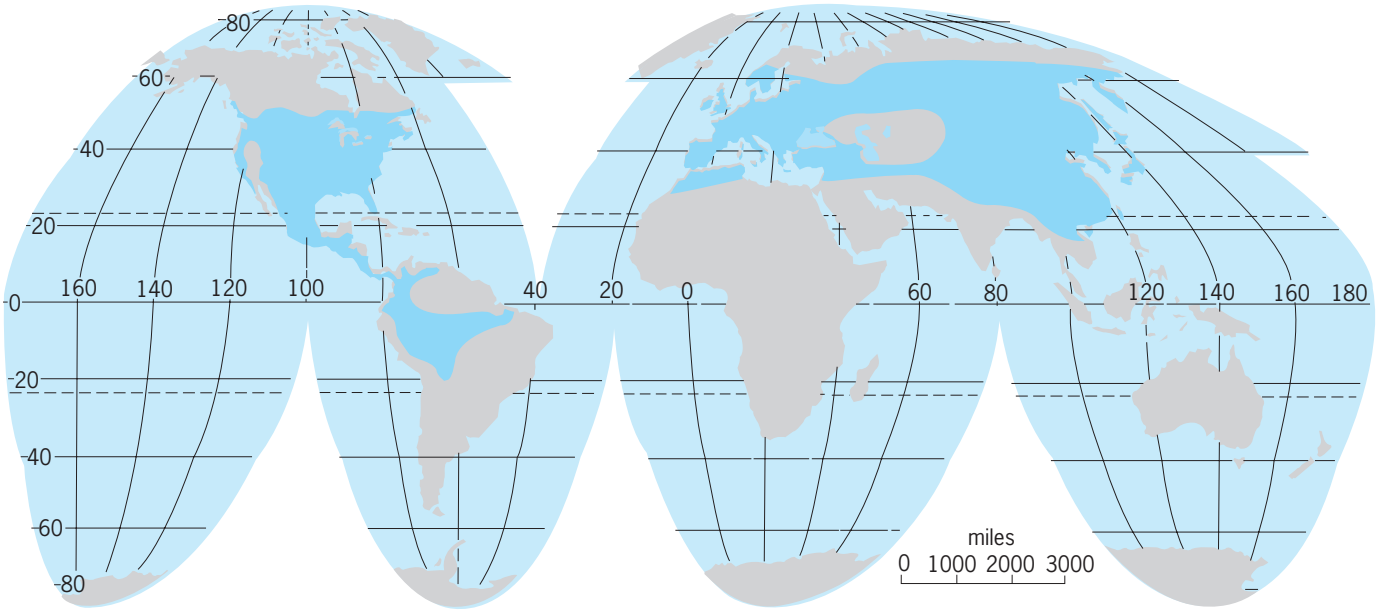
Four well-developed limbs are typically present in salamanders, but the legs of the mud eel (*Amphiuma*) of the southeastern United States are very tiny appendages and of no use in locomotion. The sirens (*Siren* and *Pseudobranchus*), also aquatic salamanders of the same region, have undergone even further degeneration of the limbs and retain only tiny forelimbs. Salamanders with normal limbs usually have four toes on the front feet and five on the rear, though a few have only four on all feet. A feature of the Urodela not shared with the Anura is the ability to regenerate limbs. If a leg is bitten off or otherwise lost, a salamander can grow a whole new limb.

Physiology. The moist and highly vascularized skin of a salamander serves as an organ of respiration, and in some forms shares with the buccal region virtually the whole burden. In one entire family of salamanders, the Plethodontidae, lungs are not present. Even in those forms with lungs and gills, dermal respiration plays a very important part.

The tympanum is absent in salamanders and the middle ear is degenerate. Hence these animals are deaf to airborne sounds. Undoubtedly correlated with this is the reduction of voice. In contrast to the highly vocal frogs, salamanders are mute or produce only slight squeaking or, rarely, barking sounds when annoyed. However, salamanders can detect sounds carried through ground or water.

Ecology. Salamanders live in a variety of aquatic or moist habitats. Many forms are wholly aquatic, living in streams, rivers, lakes, and ponds. Others live on land most of the year but must return to water to breed. The most advanced species live out their lives in moist places on land or, in the case of some tropical species, in trees, and never go in the water.

Reproduction. Producing the young alive rather than by laying eggs is very rare among salamanders. The eggs of salamanders may be deposited in water or in moist places on land. Presumably, aquatic breeding is the more primitive. Fertilization is external in most of the Cryptobranchoidea, internal by means of spermatophores in the remaining suborders. The spermatophore is a gelatinous capsule of sperm deposited by the male salamander in the course of an often elaborate courtship. The female salamander picks up the capsule with her cloacal lips and retains the sperm in her body until the eggs are fertilized and laid. The young salamander that hatches from the egg may have the adult morphology in the case of advanced forms, but usually it emerges as a larva. The larva of a salamander resembles the adult closely, in contrast to the larva of a frog, but differs in possessing gills and in being sexually immature, as well as in histological and other internal characters. There are salamanders in which metamorphosis from larval to adult morphology does not take place, and sexual maturity is attained in the larval condition. This condition is known as neoteny. In some forms it is a permanent, unchangeable condition, while in others the larva can be made to undergo metamorphosis



World distribution of salamanders (Urodela), indicated by colored areas of land. 1 mi = 1.6 km.

by treatment with thyroid hormone. *See* NEOTENY; THYROID GLAND.

Nutrition. Both as larva and as adults, salamanders are almost wholly carnivorous. Few species take anything but invertebrates in their diet, although larger forms may eat small vertebrates such as frogs and fish. Some are known to eat both plant material and invertebrates.

Distribution. The greatest concentration of families and genera of salamanders is in the eastern United States, and these animals are one of the few major groups of vertebrate animals that is distinctly nontropical. No salamander occurs below the Equator in the Old World, and only 18 species of the evolutionarily advanced family Plethodontidae are found as far south as South America in the New World (see **illus.**).

Classification. The Urodela are readily arranged in seven families, but authorities differ as to the relationships among these families and their classification into suborders. Three to five suborders may be recognized, and one of these is thought by some to merit ordinal rank, equivalent to the Urodela. About 300 species of salamanders are known to be living today, and these are distributed among about 54 genera.

Cryptobranchoidea. The Asiatic salamanders (family Hynobiidae) and the giant salamanders (family Cryptobranchidae) are sometimes grouped as the suborder Cryptobranchoidea. Five genera are recognized in the Hynobiidae, the most important of which is *Hynobius* with 18 species found over much of Asia from Japan to the Ural Mountains. One species even crosses the Arctic Circle, the farthest north reached by any salamander. There are only two living genera, each with a single species, in the Cryptobranchidae: the hellbenders (*Cryptobranchus*) of the eastern United States and the giant salamanders (*Megalobatrachus*) of Japan and China. *Andrias*, a fossil genus of Europe and North America, is apparently identical with *Megalobatrachus*, and if this is the case, the Asiatic giant salamander should be called *Andrias*.

The members of this primitive suborder differ from other salamanders in having external fertilization, though *Ranadon* evidently is an exception. Aquatic larvae are present in all forms, and some hynobiid larvae are peculiar in having sharp, recurved claws. Hynobiids undergo normal metamorphosis, but cryptobranchids retain many larval characteristics and are partially neotenic.

The hellbender and its giant Asiatic relative are aquatic species that live in rivers and streams. In the Hynobiidae there are both terrestrial and aquatic species.

Salamandroidea. This is the largest suborder of the Urodela with some 260 species in 5 families and 48 genera. All living salamanders belong to this group except the members of the suborder Cryptobranchoidea, which have external fertilization and lack the fusion of the angular and prearticular bones of the lower jaw seen in higher forms, and the suborder Meantes, neotenic forms lacking hindlimbs.

Ambystomatidae. The family Ambystomatidae, sometimes considered as a separate suborder, Ambystomoidea, is composed of 4 genera and 35 species limited in distribution to North America. The genus *Ambystoma*, with about 30 species, is the largest and most widespread of the family, and ranges from southern Alaska to Mexico and from the Atlantic to the Pacific. The tiger salamander (*Ambystoma tigrinum*) is found in a variety of subspecific forms from southern Canada to Mexico and over most of the United States. It lives in both arid and humid regions and is the only salamander in much of the region of the Great Plains and the Rocky Moun-

tains. Heavily forested regions of western North America are the home of the Pacific giant salamander (*Dicamptodon*). This largest of terrestrial salamanders may attain a length of 1 ft (0.3 m). All ambystomatids have aquatic larvae, but adults of most species are terrestrial, returning to the water only to breed. However, neoteny is of frequent occurrence in the family. The axolotl (*A. mexicanum*) of Mexico, which is neotenic in the natural state but will transform in captivity, is an ambystomatid.

Salamandridae. This family of North America, Europe, and Asia includes 40 species in 16 genera. The majority of forms are Eurasian, with only 2 genera and 6 species found in North America: *Taricha* on the Pacific Coast and *Notophthalmus* in the eastern United States and northeastern Mexico. These genera are endemic to North America, although an earlier, simplified classification placed several American and Eurasian species in a single genus, *Triturus*. The newts, as these forms are called, are often strikingly colored animals with brilliant red or orange underparts and darker brown or green upper surfaces. All newts have an aquatic larval state, whereas the adults are variously aquatic or terrestrial, according to species. A complicated life history is shown by the red-spotted newt of eastern North America, in which there is a land-dwelling stage (the red eft) commonly but not always interposed between the aquatic larval and the aquatic adult stages. Although most salamanders are nocturnal, the red eft and adults of *Taricha* and *Notophthalmus* are often found in daylight. These animals seem to be almost immune to attack by predators, possibly because of noxious skin secretions.

Two salamanders of the genus *Salamandra* are two European species with breeding habits that are peculiar for salamanders. *Salamandra atra*, a high-mountain form, retains the eggs and larvae within the body until they emerge as fully transformed young. *Salamandra salamandra*, a species of lower elevations, gives birth to larvae. In both species highly developed gills enable the larvae to absorb oxygen from the maternal oviduct.

Amphiumidae. Three large, eellike salamanders of the southeastern United States are the only members of the family Amphiumidae. These congo eels (*Amphiuma*) are partly neotenic aquatic animals with very tiny limbs of no use in locomotion.

Plethodontidae. This large and diverse family of about 180 species and 24 genera is characterized by the absence of lungs and the presence of a fine groove from nostril to upper lip. Except for two species of *Hydromantes* in Europe, the family is wholly American. Three species of this genus are found in California, but the greatest concentration of genera and species is in the eastern United States, with a secondary center in Mexico and Central America. The species found in South America are the world's southernmost salamanders.

The lungs are reduced or absent in some salamanders of other families, but this feature is consistent throughout the Plethodontidae. It is thought that the absence of lungs serves a hydrostatic function in stream-dwelling salamanders, and for this reason plethodontids are thought to have evolved from ancestors with such habits. However, the present-day members of this family are an ecologically diversified group. A majority of the species are terrestrial and lay eggs that develop directly into salamanders with adult morphology, skipping the aquatic larval stage. Other forms have aquatic larvae and terrestrial or aquatic adults. Many tropical species are arboreal, living in air plants, bromeliads, that retain moisture in dry seasons. At the other extreme are neotenic, subterranean species that come to light only when a cave is explored or a deep well dug.

Proteidae. The family Proteidae includes four species of *Necturus*, so-called mud puppies, all found in eastern North America, and the single species of *Proteus* in Europe. All are neotenic, aquatic animals. *Proteus* lives only in subterranean waters, whereas *Necturus* lives in lakes and streams. These two genera have sometimes been grouped in a separate suborder, Proteida, but it has also been suggested that they are not so closely related as they superficially appear to be.

Meantes. This group forms the smallest suborder of the Urodela, and these salamanders are sufficiently peculiar that they sometimes are regarded as constituting a separate order, Trachystomata. There are two genera, *Pseudobranchus* with a single species and *Siren* with two species. Commonly known as mud eels, these salamanders are found in areas of low elevation from Texas to Florida, Maryland, and Illinois. The mud eels are neotenic, retaining such larval characters as lidless eyes and external gills. The body form is eellike, with only anterior limbs being present. *Siren* reaches a length of 36 in. (91 cm), while *Pseudobranchus* is less than 9 in. (23 cm) long. Members of both genera are wholly aquatic, and are found beneath rocks in streams, and burrowing in the muck and vegetation of bogs and swamps. *See* AMPHIBIA; ANURA; APODA.

Richard G. Zweifel

Urogenital system

The combined structures composing the urinary and genital, or reproductive, organs of vertebrates. The terms urinogenital or genitourinary system are equally applicable and just as frequently used. During embryonic development, the urinary and reproductive systems are closely interrelated, as their ducts arise in associated mesodermal regions. Common passages are associated with both these systems in various vertebrates, such as the single orifice for emission of urine and sperm.

Urinary system. The principal organ of the urinary system is the kidney with its associated structures. Three morphologically distinct kidneys occur among the vertebrates. These are the pronephros, mesonephros, and metanephros. The pronephros is drained by the pronephric duct, the mesonephros by the Wolffian duct, and the metanephros by the

ureter. The principal function of the urinary system is the removal of nitrogenous waste products. *See* KIDNEY; URINARY SYSTEM.

Reproductive system. While the same structures of the urinary system occur in both males and females, specific structures such as the ovaries and testes are found in the normal female and male, respectively. The primary function of the reproductive system is the perpetuation of the species. The main reproductive structures of the female vertebrate consist of the ovary, oviduct, uterus, vagina, and external orifice. Many modifications occur among these structures in the various classes. Essentially, the male system consists of the testes and their ducts, seminal vesicles, prostate gland, Cowper's gland, and penis. Modifications of these structures also occur among the various vertebrate species. *See* REPRODUCTIVE SYSTEM.

Charles B. Curtin

Urokinase

An enzyme that is a plasminogen activator. Urokinase cleaves the plasma protein plasminogen, forming the active enzyme plasmin, which subsequently degrades fibrin. Thus urokinase is an essential component in the fibrinolytic clot-dissolving system in the human body. Other plasminogen activators include tissue plasminogen activator and streptokinase. *See* PLASMIN.

Urokinase is found in human urine and in much lower concentrations in human plasma. It is secreted by cells as a single-chain glycoprotein with a molecular weight of 54,000 and three structural domains: a human-growth-factor-like region, a kringle region or domain (a triple-loop protein structure containing three disulfide bonds), and a serine protease part. Unlike the tissue-type plasminogen activator produced by endothelial cells, urokinase has only a single kringle domain. The hydrolysis of urokinase by plasmin results in the formation of a two-chain molecule that possesses significantly increased activity as a plasminogen activator. *See* PROTEIN.

In the body, urokinase is produced by kidney cells, and its presence in the urine promotes the dissolution of any blood clots in the urine-collecting system of the kidneys or the bladder. Urokinase is also produced by a variety of tumor cells, and it is thought to be involved in the formation of tumor metastases. Human monocytes and fibroblasts have specific receptors for urokinase; the bound urokinase is functionally active. Urokinase can be inhibited by plasma inhibitor proteins. The presence of cell receptors and inhibitors suggests that the regulation of urokinase function is complex.

The use of urokinase to dissolve intravascular blood clots has necessitated the development of a method of production on a large scale. Initially, purified urokinase in the two-chain form was obtained by separation from human urine. However, the need for 400 gallons (1.5 kiloliters) of fresh urine to purify a single clinical dose of urokinase made that approach impractical. Subsequently, large-scale commercial production of urokinase was developed by using human embryonal kidney-cell culture. This was a major early achievement in biotechnology.

Urokinase is employed in clinical medicine in the treatment of venous blood clots (thrombophlebitis and pulmonary emboli), acute myocardial infarction, and arterial blood clots in the legs and arms. It is also used to prevent the accumulation of blood clots in intravenous catheters used to administer long-term chemotherapy. Like other plasminogen activators, urokinase is most effective in removing freshly formed blood clots. Unlike two-chain urokinase and streptokinase, tissue-type plasminogen activator binds to fibrin with high affinity; this suggests that tissue plasminogen activator is a safer and more effective thrombolytic agent. However, in clinical trials all three plasminogen activators appear to be similarly effective in removing intravascular blood clots. Consequently, the high cost of urokinase and tissue plasminogen activator, compared to streptokinase, has become the deciding factor in selecting a fibrinolytic agent. *See* BLOOD; ENZYME.

Philip C. Comp

Bibliography. F. Blasi, J.-D. Vassalli, and K. Dano, Urokinase-type plasminogen activator: Proenzyme, receptor, and inhibitors, *J. Cell. Biol.*, 104:801–804, 1987; P. M. Mannucci and A. D'Angelo (eds.), *Urokinase: Basic and Clinical Aspects*, 1982; W. J. Williams et al. (eds.), *Hematology*, 4th ed., 1990.

Uropygi

An order of arachnids, the tailed whip scorpions, comprising about 70 species from tropical and warm temperate Asia and the Americas. Most are dark reddish-brown, of medium to giant size, 0.72–2.6 in. (18–65 mm), the largest one being *Mastigoproctus giganteus* of the southern United States and Mexico. The elongate, flattened body bears in front a pair of greatly thickened, raptorial pedipalps set with many sharp spines and used to hold and crush insect prey. The first pair of legs is elongated and modified into feelers. The abdomen terminates in a slender, many jointed, whiplike flagellum. The uropygids are harmless, nocturnal creatures without poison glands that live in dark places and burrow into the soil. When disturbed, they expel a volatile liquid, with the strong odor of acetic acid, from a gland at the base of the tail. This accounts for the name "vinegaroon" given by many Americans to these much-feared animals. *See* ARACHNIDA.

Willis J. Gertsch

Uropygial gland

A relatively large, compact, bilobed, secretory organ located at the base of the tail (uropygium) of most birds having a keeled sternum. It is known also as

the preen, oil, or scent gland. This is the only true skin gland possessed by this class of vertebrates. The gland develops from a pair of invaginations of the dorsal ectoderm on each side of the free coccygeal vertebrae, giving rise to numerous secretory tubules surrounded by a sheath of connective tissue of mesodermal origin. Experiments with duck embryos have shown that formation of the gland is dependent upon a specific action of the subjacent mesoderm on the overlying ectoderm. Implanted mesoderm from other body regions failed to induce a gland-forming response in the ectoderm. *See* EMBRYONIC INDUCTION.

Wide variations in the size, shape, and structure of this gland occur among different species: it is large in most aquatic birds; rudimentary and nonfunctional in most goatsuckers, whippoorwills, nighthawks, and some pigeons; absent in bustards, many pigeons and parrots, and in all ratite birds (such as the ostrich, cassowary, and emu), although present in the embryos of all those named. *See* RATITES.

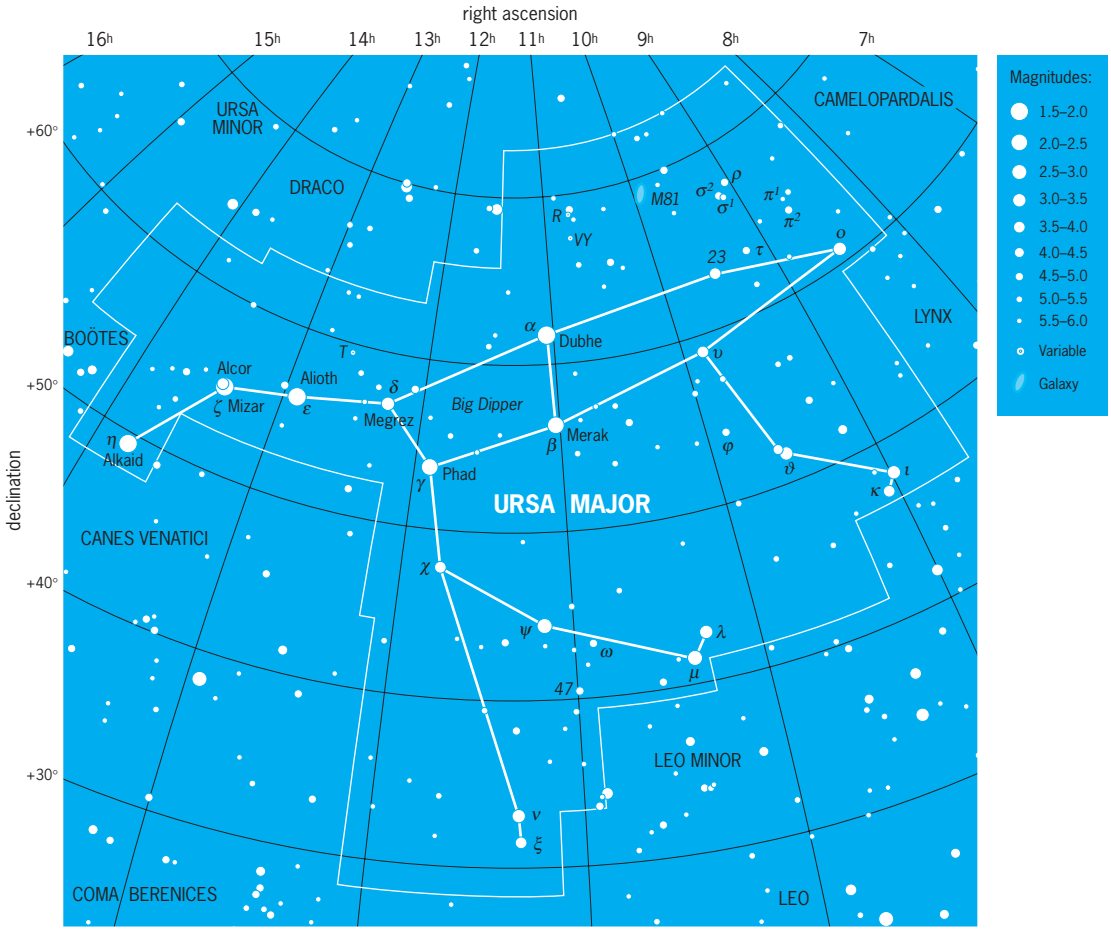
The glandular secretion, predominantly oily and sometimes of offensive odor (musk-duck, hoopoe, petrel), is discharged through an orifice at the tip of a nipplelike protuberance often encircled by short, bristly feathers. The act of preening induces a flow of secretion from the nipple which is transferred by the beak to the body plumage. In water fowl there is some evidence that this oily secretion assists in maintaining the water-repellent quality of the feathers, either directly or by preserving their physical structure. *See* FEATHER.

The occurrence and nonoccurrence of this gland in closely related species make its use difficult to interpret. Consequently, its function has aroused much controversy for over a century. No indispensable function has yet been agreed upon. It probably has different uses in different birds. *See* AVES; GLAND; SCENT GLAND.

Mary E. Rawles

Ursa Major

The Great Bear, a circumpolar constellation for observers at mid-northern latitudes or higher latitudes (see **illustration**). Part of this constellation is the prominent asterism (star pattern) known as the Big



Modern boundaries of the constellation Ursa Major, the Great Bear. The celestial equator is 0° of declination, which corresponds to celestial latitude. Right ascension corresponds to celestial longitude, with each hour of right ascension representing 15° of arc. Apparent brightness of stars is shown with dot sizes to illustrate the magnitude scale, where the brightest stars in the sky are 0th magnitude or brighter and the faintest stars that can be seen with the unaided eye at a dark site are 6th magnitude. (*Wil Tirion*)

Dipper, whose bowl of four stars is linked to a tail of three more, perhaps representing a long bear's tail pulled out when the bear was swung by it to throw it into the heavens. In England, the Big Dipper is more usually known as Charles's Wain (Wagon) or the Plough.

The stars of the Dipper were labeled by Johan Bayer (1601) in alphabetical order with Greek letters, contrary to his usual practice of assigning Greek letters in order of brightness. The two stars at the head of the dipper, alpha and beta Ursae Majoris, point toward Polaris, a star that is now within 1° of the north celestial pole. Thus these pointers are very useful for formal and informal navigation. *See* POLARIS.

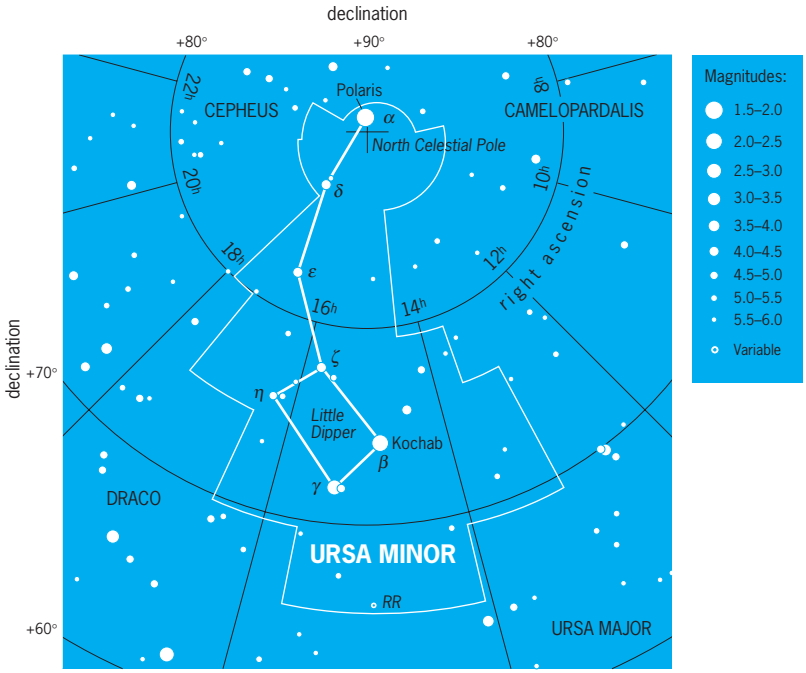
In the handle of the Big Dipper, the middle star Alcor has a fainter companion, Mizar, that is separately visible to the unaided eye. On inspection with a telescope, each is a double star. *See* MIZAR.

In Greek mythology, the bear is Callisto, a woman with whom Zeus had a son, perhaps changed in form by Zeus to protect her from the jealousy of Hera, Zeus's wife. In other versions of the myth, Callisto was changed into a bear in revenge by Callisto's companion Artemis or by Hera.

M81 is a prominent spiral galaxy in the northern part of Ursa Major, thus making it visible year-round from mid-northern latitudes. *See* GALAXY, EXTERNAL; MESSIER CATALOG.

The modern boundaries of the 88 constellations, including this one, were defined by the International Astronomical Union in 1928. *See* CONSTELLATION.

Jay M. Pasachoff



Modern boundaries of the constellation Ursa Minor, the Little Bear. The celestial equator is 0° of declination, which corresponds to celestial latitude. Right ascension corresponds to celestial longitude, with each hour of right ascension representing 15° of arc. Apparent brightness of stars is shown with dot sizes to illustrate the magnitude scale, where the brightest stars in the sky are 0th magnitude or brighter and the faintest stars that can be seen with the unaided eye at a dark site are 6th magnitude. (*Wil Tirion*)

Ursa Minor

The Little Bear, a circumpolar constellation from northern latitudes (see **illustration**). Part of this constellation is the prominent asterism (star pattern) known as the Little Dipper, whose stars are fainter than the corresponding stars of the Big Dipper. At the end of the handle of the Little Dipper is Polaris, at magnitude 2.0 not a particularly bright star—only the 49th brightest in the sky—but one that happens in the present epoch to be close to (almost 1° away from) the north celestial pole, the extension of the Earth's axis of rotation into the sky. Thus the Little Dipper and Polaris are useful for formal and informal navigation. *See* POLARIS.

In Greek mythology, Arcas, the son of Zeus and Callisto, was about to shoot a bear, not knowing that it was his mother. To protect her, Zeus changed Arcas into a bear and hurled both bears off by their tails to the heavens. Hera, annoyed at their being protected, then convinced Poseidon not to allow the bears to bathe in the sea, which resulted in their being circumpolar constellations, which never drop to the horizon. *See* URSA MAJOR.

The modern boundaries of the 88 constellations, including this one, were defined by the International Astronomical Union in 1928. *See* CONSTELLATION.

Jay M. Pasachoff

Ustilaginomycetes (smut)

A class of the subdivision Basidiomycotina. These microscopic plant-parasitic fungi are commonly known as smut fungi. Many species may have a shorter or longer saprophytic life cycle, capable of yeastlike reproduction. Several nonparasitic yeasts have been identified as closely related to the smuts. About 1500 true Ustilaginales species are known.

The body (mycelium) of the smut fungi consists of hyphae that are thin, transparent, branched, septate, and binucleate. Usually parasitic, the fungus grows in the host tissues and gives little, if any, evidence of its presence before spore formation sets in. The smut spores (teliospores or ustilospores) are the organs of dispersion and resistance, and are thick walled, pigmented, and variously ornamented. They are formed in great number in the sori, which consist of host tissues, spore masses, and sometimes modified fungal cells or tissues.

The smut fungi develop variable sori in different organs (such as the roots, stems, leaves, inflorescences, flowers, anthers, and seeds) filled by powdery or agglutinated, usually colored (black, brown, violet, or yellow) spore masses. The spores may develop singly, in pairs, or in aggregates of spore balls. These balls may be composed entirely of fertile

spores, or of a combination of spores, sterile cells, or hyphae. Spore germination results in a basidium (promycelium or ustidium) which gives rise to hyaline basidiospores. *See* BASIDIOMYCOTA; EUMYCOTA; FUNGI.

Until recently, the Ustilaginales have been divided into two families: Ustilaginaceae (with septate basidia producing lateral and apical basidiospores) and Tilletiaceae (with unseptate basidia producing apical basidiospores). In the new classification, based also on ultrastructural and molecular biological characters, the true smut fungi are divided into 2 classes, 8 orders, 17 families, and 62 genera.

Many smut fungi produce severe losses of cereals and other cultivated plants. Some important smut diseases are dwarf bunt and the two stinking or common bunts of wheat, barley, and other Gramineae (caused by *Tilletia contraversa, T. caries*, and *T. foetida*), karnal or partial bunt of wheat (caused by *Tilletia/Neovossia indica*), partial bunt of rice (caused by *Tilletia/Neovossia horrida*), corn smut (caused by *Ustilago maydis*), loose smut of wheat, barley, and oat (caused by *U. tritici, U. nuda,* and *U. avenae*), covered smut of barley (caused by *U. hordei*), sugarcane smut (caused by *U. scitaminea*), and potato smut (caused by *Thecaphora solani*). Kálmán Ványky

Bibliography. International Mycological Institute, *C.M.I. Description of Pathogenic Fungi and Bacteria*, 1964– ; K. Ványky, *Illustrated Genera of Smut Fungi*, 1987.

Uterine disorders

Neoplastic, functional, congenital, and inflammatory diseases of the uterus and cervix. Uterine and cervical disorders are common gynecological complaints, which can be categorized in terms of increased uterine size, abnormal bleeding, and infection.

Size. The most common cause of increased uterine size is the presence of fibroids (leiomyoma). These benign smooth-muscle tumors are the most common tumors in women of reproductive age, occurring in approximately one-third of women. Fibroids vary in size from less than an inch to the equivalent of a term pregnancy. Problems arise when the fibroid presses on nearby organs such as the urinary tract. Fibroids may also cause infertility or pregnancy loss. Rarely, fibroids become malignant. With any increase in uterine size, pregnancy must be excluded and medical attention sought. Large fibroids must be surgically removed (myomectomy), and the uterus may have to be removed as well (hysterectomy). Small fibroids may be treated medically by turning off ovulation, which lowers estrogen levels and shrinks the tumors. Fibroids usually become smaller after menopause.

Abnormal bleeding. Abnormal uterine bleeding, especially in postmenopausal women, must always be evaluated. In girls and young women it is usually caused by irregular ovulation or hormonal imbalances. In women of reproductive age, problems of pregnancy must always be excluded such as miscarriage or ectopic pregnancy. Other common causes include polyps of the uterus or cervix. These benign, fleshy tumors should be removed to exclude cancer, but are usually of no consequence. Occasionally, the uterine lining (endometrium) invades the muscle of the uterus, making the uterus less able to compress the surrounding arteries during menstruation. This condition, known as adenomyosis, typically causes heavy, painful menses. *See* MENSTRUATION.

The most serious reason to seek medical attention for abnormal uterine bleeding is to exclude the possibility of cancer. Uterine and cervical cancers are the most common cancers of the female genital system. Uterine cancers usually occur in women who are obese, are hypertensive, and have diabetes mellitus; cervical cancer is most common among women who have engaged in sexual intercourse at an early age with multiple partners, especially uncircumsized ones. The diagnosis of uterine cancer is made by an endometrial biopsy or dilatation and curettage. The cornerstone of diagnosing cervical cancer is the Pap smear, which examines the cells of the cervical canal. Noninvasive or minimally invasive cancers may be treated by surgery alone. More advanced cancer requires either radiation or chemotherapy as well.

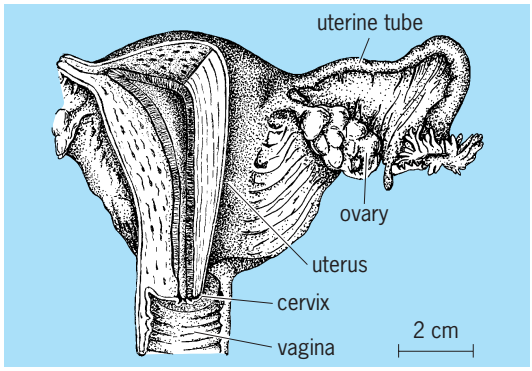
Infection. Uterine infections are not extremely common but are associated with trauma, childbirth, or miscarriage. The inflammation may be associated with a foul discharge, pain, or bleeding. Cervical infections and inflammations (cervicitis) are common, and symptoms include a foul discharge and staining or bleeding. Most inflammations are nonspecific but may become chronic. Specific agents of cervicitis include staphylococci, streptococci, gonococci, syphilis, chancroid, tuberculosis, chlamydia, and mycoplasma; appropriate antibiotics should be chosen to treat such infections. *See* REPRODUCTIVE SYSTEM DISORDERS; STAPHYLOCOCCUS; STREPTOCOCCUS; SYPHILIS; TUBERCULOSIS; UTERUS. Machelle M. Seibel

Bibliography. H. Melchoir, D. Walwiener, and G. Schlag (eds.), *Gynecology and Obstetrics*, 1994.

Uterus

The hollow, muscular womb, being an enlarged portion of the oviduct in the adult female. An adult human uterus, before pregnancy, measures approximately $3 \times 2 \times 1$ in. ($8 \times 5 \times 3$ cm) in size and has the shape of an inverted, flattened pear. The paired Fallopian tubes enter the uterus at its upper corners; the lower, narrowed portion, which is called the cervix, projects into the vagina (see **illus.**). Normally the uterus is tilted slightly forward and lies behind the urinary bladder.

The lining, or mucosa, responds to hormonal stimulation, growing in thickness with a tremendous increase in blood vessels during the first part of the menstrual cycle. If fertilization does not occur, the thickened vascular lining is sloughed off, producing the menstrual flow at the end of the cycle, and a new



Human uterus and associated structures. (*After L. B. Arey, Developmental Anatomy, 7th ed., Saunders, 1965*)

menstrual cycle begins with growth of the mucosa. When pregnancy occurs, the mucosa continues to thicken and forms an intimate connection with the implanted and enlarging placenta. However, the maternal and fetal blood vessels remain distinct; food, oxygen, and wastes flow between these two circulatory systems, although maternal and fetal bloods do not mix. *See* MENSTRUATION; PREGNANCY.

The uterus develops from two embryonic tubes, which may remain distinct in the adult, forming two uterine chambers as in the rabbit; or may fuse to various degrees, forming a branched two-chambered uterus with a single lower portion, as in the cow; or may unite to form a single chamber, as in humans. *See* REPRODUCTIVE SYSTEM.

Walter Bock

Vaccination

Active immunization against a variety of microorganisms or their components, with the ultimate goal of protecting the host against subsequent challenge by the naturally occurring infectious agent.

The terms "vaccine" and "vaccination" were originally used only in connection with Edward Jenner's method for preventing smallpox, introduced in 1796. In 1881 Louis Pasteur proposed that these terms should be used to describe any prophylactic immunization. Vaccination now refers to active immunization against a variety of bacteria, viruses, and parasites (for example, malaria and trypanosomes). [**Tables 1** and **2**].

During the last 200 years, vaccination has controlled nine major diseases: smallpox, diphtheria, tetanus, yellow fever, pertussis, poliomyelitis, measles, mumps, and rubella. In the United States and other developed countries, where programs have been expanded to immunize virtually all children under 2 years of age, most vaccine-preventable diseases have been reduced by 95% or more from prevaccine highs (**Table 3**). In the case of smallpox, the dream of eradication has been fulfilled. Vaccinations against influenza, hepatitis B, pneumococci, and hemophilus influenza type b have made major headway against these infections. *See* SMALLPOX.

Implicit within Jenner's method of vaccinating against smallpox was the recognition of immunologic cross-reactivity together with the notion that protection can be obtained through active immunization with a different, but related, live virus. It was not until the 1880s that the next immunizing agents, vaccines against rabies and anthrax, were introduced by Pasteur. Two facts of his experiments on rabies vaccines are particularly noteworthy.

First, Pasteur found that serial passage of the rabies agent in rabbits resulted in a weakening of its virulence in dogs. For the next 100 years, Pasteur's empirical approach for attenuating the virulence of a live virus by repeated passages in cells of species different from the natural host remained the principal empirical method for developing attenuated-live-virus vaccines. During multiple passages in an animal or in tissue culture cells, mutations accumulate as the virus adapts to its new environment. These mutations adversely affect virus reproduction in the natural host, resulting in lessened virulence. Only as the molecular basis for virulence has begun to be elucidated by modern biologists has it become possible to deliberately remove the genes

TABLE 1. Some bacterial vaccines licensed in the United States

Vaccine	Year first described	Vaccine type
Typhoid	1896	Inactivated
Cholera	1896	Inactivated
Plague	1897	Inactivated
Pertussis	1926	Acellular toxoid
Bacillus Calmette-Guérin (tuberculosis)	1927	Live attenuated
Diphtheria	1888	Toxoid (protein)
Tetanus	1888	Toxoid (protein)
Pneumococcus Meningococcus Hemophilus influenza b	Post World War II	Capsular polysaccharide

TABLE 2. Some vaccines used in prevention of viral diseases of humans

Disease	Condition of virus	Route of administration	Comment
Poliomyelitis (Sabin)	Live attenuated	Oral	Recommended immunization for general human population
Poliomyelitis (Salk)	Inactivated	Subcutaneous	
Measles	Live attenuated	Subcutaneous	
Rubella*	Live attenuated	Subcutaneous	
Mumps	Live attenuated	Subcutaneous	
Mumps (Finland)	Inactivated	Subcutaneous	
Hepatitis A†	Inactivated	Intramuscular	Immunization recommended only during epidemic, travel, military service, exposure, or for selected high-risk groups
Smallpox	Vaccinia virus, an apothogenic virus related to variola	Intradermal	
Yellow fever	Live attenuated	Subcutaneous	
Influenza	Live attenuated	Oropharyngeal	
Influenza	Inactivated	Subcutaneous	
Rabies	Inactivated	Subcutaneous	
Adenovirus infection	Live attenuated	Oral by enteric coated capsules	
Japanese B encephalitis	Inactivated	Subcutaneous	
Eastern and western equine encephalitis	Inactivated	Subcutaneous	
Chickenpox‡	Live attenuated	Subcutaneous	
Hepatitis B (Heptavax-B)§	Subunit (HBsAg)	Subcutaneous	
Hepatitis B (Institut Pasteur)¶	Subunit (HBsAg) including Pre-S peptide	Subcutaneous	
Hepatitis B	Subunit (recombinant DNA from yeast)	Subcutaneous	
Cytomegalovirus disease	Live attenuated	Subcutaneous	Experimental vaccine being tested on pediatric nurses and renal transplant candidates

*Not to be administered to a postpubertal woman unless she is not pregnant and understands that it is imperative she not become so for at least 3 months after vaccination.

†Available for children over 2 years of age and for adults. Recommended for travelers to regions of high endemic disease, children living in countries with high prevalence, health-care workers, users of illicit intravenous drugs, food handlers, toddlers in day care and their supervisors, recipients of liver transplants, and persons older than 30 years with chronic liver disease.

‡Used for leukemic children and other high-risk patients.

§High-risk groups in the United States are health-care workers, selected patients on hemodialysis, children with hemophilia, residents and staff of institutes for the mentally handicapped, homosexually active males, prostitutes, users of illicit drugs, prisoners, certain military personnel, and infants and young children immigrating from high epidemic areas.

¶Used in Taiwan in 1985 to immunize all newborn babies who were at risk for infection from their mothers.

promoting virulence so as to produce attenuated viruses.

Second, Pasteur demonstrated that rabies virus retained immunogenicity even after its infectivity was inactivated by formalin and other chemicals, thereby providing the paradigm for one class of noninfectious virus vaccine, the "killed"-virus vaccine.

Attenuated-live and inactivated vaccines are the two broad classifications for vaccines. Anti-idiotype antibody vaccines and deoxyribonucleic acid (DNA) vaccines represent innovations in inactivated vaccines. Recombinant-hybrid viruses are novel members of the live-virus vaccine class recently produced by genetic engineering.

Attenuated-live vaccines. Because attenuated-live-virus vaccines reproduce in the recipient, they provoke both a broader and more intense range of antibodies and T-lymphocyte-associated immune responses than noninfectious vaccines. Live-virus vaccines have been administered subdermally (vaccinia), subcutaneously (measles), intramuscularly (pseudorabies virus), intranasally (infectious bovine

TABLE 3. Past maximum and current morbidity from vaccine preventable diseases

Disease	Number of cases	
	Maximum (year)	In 1998
Diphtheria	206,939 (1921)	1
Measles	894,134 (1941)	100
Mumps	152,209 (1968)	666
Pertussis	265,269 (1934)	7405
Polio (paralytic)	21,269 (1952)	1
Rubella	57,686 (1969)	364
Congenital rubella syndrome	20,000 (1964-65)	7
Tetanus	1,560 (1923)*	41

*Deaths in death registration states only in 1923.

rhinotracheitis), orally (trivalent Sabin poliovirus), or by oropharyngeal aerosols (influenza). Combinations of vaccines have also been used. When combined live measles, rubella, and mumps vaccines have been given by injection, the antibody response to each component has been comparable with the antibody response to the individual vaccines given separately.

Many organisms produce their pathophysiologic effects through acute infections localized to the gastrointestinal, pulmonary, nasopharyngeal, and genitourinary surfaces. These areas are bathed in mucus that contains highly important secretory immunoglobulin A (IgA). Subcutaneous, intravenous, and intramuscular immunization regimens, though effective in inducing IgG-type antibodies, are almost universally ineffective at the induction of secretory IgA antibodies. In contrast, oral administration is a feasible alternative to stimulate the common mucosal immune system, as with the live-attenuated polio vaccine. Live-virus vaccines administered through a natural route of infection often induce local immunity, which is a decided advantage. However, in the past, attenuated-live virus vaccines have been associated with several problems, such as reversion to virulence, natural spread to contacts, contaminating viruses, lability, and viral interference. *See* IMMUNOGLOBULIN; VIRULENCE; VIRUS CLASSIFICATION; VIRUS INTERFERENCE.

Noninfectious vaccines. Noninfectious vaccines include inactivated killed vaccines, subunit vaccines, synthetic peptide and biosynthetic polypeptide vaccines, oral transgenic plant vaccines, anti-idiotype antibody vaccines, DNA vaccines, and polysaccharide-protein conjugate vaccines. With most noninfectious vaccines a suitable formulation is essential to provide the optimal antigen delivery for maximal stimulation of protective immune responses. Development of new adjuvant and vector systems is pivotal to produce practical molecular vaccines.

Killed vaccines. The inactivated forms of the toxins produced by the tetanus and diphtheria bacteria, called toxoids, were among the first vaccines developed. For human use, toxoid vaccines must be highly purified to remove extraneous materials. Killed-virus vaccines consist of partially purified virus particles whose infectivity has been destroyed by treatment with chemicals or with radiation. These virus particles retain the ability to elicit an immune response and to protect from clinical disease. Though generally safer and more stable than attenuated-live-virus vaccines, killed-virus vaccines require multiple doses of high concentrations of antigen sometimes administered with adjuvant, that is, a substance that enhances the potency of the antigen. The induction of cellular immune responses may be poor and the protection induced by killed-virus vaccines may be of short duration, so that booster vaccinations are needed. Vaccines produced by killing viruses have been known to contain surviving, infectious particles that have caused outbreaks of disease (Cutter polio vaccine, foot-and-mouth-disease vaccine). Although incapable of producing direct cellular destruction, noninfectious vaccines may elicit abnormal hypersensitivity reactions.

Through the use of the oral, attenuated-live (Sabin) vaccine and the inactivated (Salk) vaccine, paralytic polio has been virtually eliminated from the United States. The last naturally occurring case of paralytic polio in the United States was in 1979. In 1997 and 1998, there were only six reported cases. All were associated with receipt of the trivalent oral, attenuated-live vaccine. An all-inactivated polio vaccine schedule has been recommended for routine childhood vaccination beginning January 1, 2000. *See* CELLULAR IMMUNOLOGY; DIPHTHERIA; POLIOMYELITIS; TETANUS.

In the 1970s, concerns developed that whole-cell pertussis vaccines caused permanent brain damage in some children. This led to a sharp decline in pertussis vaccination, waning immunity, and outbreaks of disease in several countries. Even though the general risk of serious postvaccination neurologic damage was not confirmed, a strong feeling persisted that safer and more efficacious pertussis vaccines were needed. Acellular pertussis vaccines consisting of pertussis toxoid (and in some cases, highly purified filamentous hemagglutinin and pertactin proteins) were recommended in the United Kingdom, Sweden, and Japan, and were approved in 1996 in the United States. *See* WHOOPING COUGH.

Polysaccharide vaccines. Polysaccharide vaccines consist of the polysaccharide coats, or capsules, of encapsulated bacteria. Although polysaccharides are recognized mainly by T-cell-independent mechanisms, they do not effectively produce high-level, high-affinity antibodies, nor do they induce the T-cell memory required for booster responses. However, the immunogenicity of polysaccharides can be significantly increased by covalently linking them to carrier proteins. This approach is thought to work through the recruitment of T-cell help. For example, Hemophilus influenza b vaccines developed with this conjugate approach have been very successful in inducing high, boostable, protective antibody levels in infants and have dramatically reduced Hemophilus influenza b–associated disease in countries where they are in general use. *See* INFLUENZA.

Subunit vaccines. Subunit vaccines consist of immunogenic viral proteins stripped free from whole virus particles, then purified from other irrelevant components, thereby reducing the risk of adverse reactions and residual infectious virus. Subunit vaccines require the addition of adjuvants or their formation into forms such as lipid vesicles (liposomes, virosomes) or micelles. The inability to propagate hepatitis B virus in tissue culture provided an incentive for the production of one of the first subunit vaccines, Heptavax-B, which was obtained by purifying and formalin-inactivating the surface antigen from the plasma of chronic carriers of hepatitis B. Experimental subunit vaccines have also been produced from purified glycoproteins derived from

the virus membranes of enveloped virus particles, such as parainfluenza-3, measles, rabies, influenza, respiratory syncytial virus (RSV), feline leukemia virus, bovine herpes virus-1, and pseudorabies. *See* LIPOSOMES.

Immune stimulating complexes (ISCOMs) are very immunogenic and represent an interesting delivery system for glycoprotein subunit vaccines. These complexes are really advanced liposomes that look like small balls and are made from an adjuvant.

Synthetic peptides. Chemically synthesized peptides are perceived as one alternative to conventional vaccines to elicit virus-neutralizing antibodies, or to "prime" the host for neutralizing antibody responses. A conventional approach to preparing antipeptide antibodies is conjugation of a peptide to a known protein or synthetic polymer carrier. Methods designed to avoid the use of carrier by polymerizing or cyclizing the peptides have also been reported, as has an approach known as the multiple-antigen peptide system, in which a small peptidyl core matrix is covalently bound to radially branching synthetic peptides as dendritic arms.

The peptide sequences chosen for synthesis in developing new vaccines correspond to antigenic determinants (epitopes) on the surface of virus particles that have been identified through chemical, electron-microscopic, and crystallographic analysis. Application of recombinant DNA technology has also allowed nucleotide sequences of viral genomes to be determined, from which primary amino acid sequences, secondary structure, and potential B-cell antigenic and T-cell (amphipathic) sequences may be deduced.

Carrier-free polypeptide vaccines are anticipated to be especially safe because they are chemically well defined, do not contain nucleic acids or extraneous proteins, and are unlikely to have any of the pathogens that might be present in serum or tissue extracts. Polypeptide vaccines are also extremely stable, withstanding thermal extremes and ambient conditions that might inactivate other types of vaccines under field conditions. This is important where target populations in underdeveloped countries may lack access to refrigeration equipment.

Synthetic peptide vaccines may be feasible for agents for which protection is based mainly on humoral (antibody-mediated) immunity and suitable conserved antigenic sites can be found. Synthetic peptide vaccines are more realistic for DNA than for ribonucleic acid (RNA) viruses, since DNA viruses mutate much less frequently than RNA viruses. Thus, antibodies against a single antigenic site of an RNA virus may allow too many escape mutants and, in contrast to DNA viruses, may require antibodies against different antigenic sites. Successful synthetic peptide vaccines have been produced for the simple, DNA-containing canine parvovirus and mink enteritis virus, which readily induce full protection of all vaccinated animals after a severe challenge with virulent viruses. *See* CANINE PARVOVIRUS INFECTION; FELINE PANLEUKOPENIA; PEPTIDE.

Biosynthetic polypeptide vaccines. The nucleic acid sequences encoding microbial antigens and chemically synthesized peptides can be cloned in bacterial and yeast plasmids or in viral vectors and expressed efficiently in prokaryotic and eukaryotic cells, either as genetically purified (cloned) proteins or as parts of fusion proteins. The immunogens made in this way may be secreted into the medium or may be expressed on the surface or in the cytoplasm of cultured cells. The antigens can then be purified from the culture medium or from cell lysates. The biosynthetic production of antigens is particularly useful when the microbial agent is difficult to grow in tissue culture, such as human hepatitis B virus, or dangerous, such as rabies or HIV.

Anti-idiotype antibody vaccines. The binding site of an antibody possesses a highly distinctive three-dimensional structure (the idiotype) that can itself act as an antigenic determinant to stimulate the formation of antibodies. The anti-idiotype antibodies, therefore, mimic the form of the surface antigen initially employed to induce the formation of a specific idiotypic antibody.

To prepare such vaccines, a monoclonal antibody is first produced specifically against a microbial antigen. This monoclonal antibody is injected into an animal to induce a second antibody specific to the monoclonal antibody. If the second antibody binds to the idiotype of the initial monoclonal antibody, then the antigen-recognition portion of the second antibody should be structurally analogous to the original microbial neutralization antigen. Injection of the second antibody as a vaccine will then induce antibodies against the original microbial agent. Experimental anti-idiotype antibody vaccines have been made against a number of diseases including hepatitis B virus, rabies virus, HIV, herpesvirus, poliovirus, and several bacterial and parasitic pathogens. *See* ANTIBODY.

DNA vaccines. DNA (gene) vaccines, or the direct injection of DNA plasmids that express antigens of interest, represent another approach to subunit vaccines. In animal models, DNA vaccines stimulate cell-mediated as well as antibody-based immune responses, unlike most conventional inactivated vaccines. The DNA vaccines persistently express genes encoded by the plasmid without integrating into chromosomal DNA. Gene expression from the direct intramuscular injection of DNA has been detected in mice, rats, rabbits, chicks, dogs, fish, cattle, and nonhuman primates, illustrating the diversity of species in which direct DNA vaccination without a delivery vehicle is possible. DNA vaccines have a number of advantageous properties. They can protect against viral infections in which the antibody response alone is not protective, or where there is pronounced antigenic diversity among the target strains. Protection against lethal challenges has been achieved through parenteral or mucosal immunizations, or by the use of gene guns that deliver tiny amounts of DNA-coated gold beads. The same or similar DNA plasmids can be used as vectors for subsequent immunizations, because no immune response is elicited

to the vector. The development of new DNA vaccines may be pertinent to several diseases, including acquired immune deficiency syndrome (AIDS) and hepatitis B, and will undoubtedly receive increasing attention. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS).

Adjuvants. Adjuvants are chemicals that significantly enhance speed, vigor, and persistence of strong antigens and the potency of weak antigens. Adjuvant formulations consist of adjuvants in suitable delivery vehicles, such as mineral or vegetable oil emulsions, squalene, liposomes, nonionic block polymer surfactants, and biodegradable polymer microspheres. Biodegradable microspheres can be especially useful as vehicles because of the ease of vaccine delivery by the oral route and the safety of the delivery vehicle (tissue compatibility and no secondary reactions). The most common microspheres range from a few micrometers to 200 μm in diameter.

The most common adjuvants used in immunological research are Freund's incomplete adjuvant, which consists of an aqueous antigen solution emulsified with an equal volume of paraffin oil/mannide monooleate (Arlacel A), and Freund's complete adjuvant (muramyl dipeptide delivered in oil emulsions), which also contains killed mycobacteria. The complete adjuvant consistently stimulates high and long-lasting antibody responses, which can be attributed to the slow release of antigen from the site of injection. Neither form of adjuvant is acceptable for human use because both induce pain, abscess formation, granulomas, synovial joint lesions, and other adverse reactions. The only adjuvants authorized in the United States for human use are alum salts, which are both ineffective with some antigens and capable of producing adverse reactions, such as granulomas at the injection site. The emergence of subunit vaccines created by recombinant DNA technology and of synthetic peptide vaccines has intensified the need for safer and more effective adjuvants.

Cytokines and immunomodulation. The induction of an immune response in the presence of an appropriate antigen initiates reciprocal interactions between antigen presenting cells, such as dendritic and Langerhans cells (and possibly macrophage) and T lymphocytes bearing receptors for processed forms of the antigens. The development and propagation of the immune response also involves diverse subtypes of T cells and B lymphocytes, natural killer and lymphokine activated killer cells, neutrophils, and eosinophils. The communications between these immune and inflammatory cells are mediated in large part by small proteins (about 20 kilodaltons) named cytokines. The cytokines include the interleukins (IL-1 to IL-18), gamma interferon, granulocyte-macrophage colony stimulating factor (GM-CSF), tumor necrosis factor, and transforming growth factor beta. The cytokines have various actions, for example, IL-1 (T- and B-cell maturation), gamma interferon [helper T cell (Th)1 upregulation, enhanced major histocompatibility complex expression], IL-2 (Th1 upregulation), IL-4 (Th2 upregulation), IL-10 (suppression of inflammatory cytokines, enhancement of B-cell proliferation), GM-CSF (co-migratory signals for dendritic cells), and IL-12 (stimulation of Th1 differentiation). Cytokines are species-specific and expensive, and there are concerns regarding their stability, toxicity, and potential autoimmunity. These issues have limited their use as vaccine components. *See* CYTOKINE; IMMUNOLOGY.

Immunomodulation refers to the ability of many adjuvants to modify the cytokine network and, hence, the immune response. Adjuvants can target macrophages of the Peyer's patches (lymph nodules of the intestine) to induce mucosal immune responses, or allow antigens to associate directly with Class I major histocompatibility complex (MHC) molecules on the cell surface to elicit Class I–restricted CD8+ cytotoxic T-lymphocyte responses. Particulate antigens tend to target macrophages, and macrophages can be recruited to the site of antigen deposition through the action of cytokines. Freund's complete adjuvant and lipopolysaccharide-based adjuvants drive the responses of Th1 subtype CD4+ helper cells, while alum salt, cholera toxin, and pertussis adjuvants drive the responses of Th2 subtype CD4+ helper cells. The Th1 responses are associated with interferon gamma, IL-2, and IL-12 secretions and T cells involved with strong delayed-type hypersensitivity reactions. Th1 cells also secrete tumor necrosis factor, thereby enhancing B-cell responses that elicit IgG2a, the antibody subclass most efficient in binding the serum complement proteins that enhance antigen-antibody reactions. The Th2 cells secrete the IL-4, IL-5, and IL-10 that activate the production of high levels of IgG1, IgA, and IgE by B cells. Selection of the appropriate immunoregulatory adjuvant not only leads to an enhanced immune response but also determines the isotype of IgG, which other immunoglobulins are made, and how much CD4+-directed cell-mediated immunity is generated. *See* IMMUNOLOGIC CYTOTOXICITY.

Recombinant viral vectors and chimeric viruses. Another option is the use of recombinant viruses for the delivery of vaccine antigens. Although foreign antigens may be expressed in these recombinant viruses and used as inactivated subunit vaccines, live-attenuated-viral-based vectors as vaccines have decided advantages. This is because a single vaccination without adjuvants can suffice for immunization by the live-virus vaccine. Replication of the viral-based vector within the host amplifies the amount of foreign antigen being expressed, thereby often increasing both cell-mediated and humoral immune responses to the foreign antigen. The processing and posttranslational modification of the foreign antigen within infected eukaryotic cells more closely resemble those occurring during natural infection with the pathogen from which the foreign antigen was derived. In addition, presentation of the antigen may be optimized.

Attenuated bacteria, such as *Salmonella*, as well as many different virus families have been used as vectors, including adenoviruses, herpesviruses,

papovaviruses, retroviruses, and picornaviruses. However, the most extensively developed and exploited vector has been vaccinia virus, a member of the Poxvirus family. Vaccinia is a very large virus; about 5% of its DNA either is redundant or encodes genes that are not absolutely essential for virus replication in cultured cells. Thus, the capacity of vaccinia recombinants to accommodate foreign DNA is very large. Experimental vaccinia-based-vector vaccines include those with foreign DNA inserts derived from hepatitis B virus, influenza, malaria, Epstein-Barr virus, genital herpesvirus, respiratory syncytial virus, Lassa fever virus, human T-cell leukemia virus, and AIDS virus. One of the most interesting of the vaccinia-based viral vectors is the recombinant that expresses the rabies virus glycoprotein gene. The vaccinia/rabies recombinant was incorporated in bait as an oral vaccine and is presently being used for the immunization of wild life. Although vaccinia is a comparatively safe vaccine virus, immunosuppressed people can become systemically infected and die with generalized vaccinia infection. Therefore, it is difficult to imagine vaccinia being used in humans in view of the global AIDS epidemic. An additional problem arises from the induction of vaccinia antibodies during the initial usage of the vaccinia vector so that it becomes difficult to use the vectorial vaccine twice.

Recombinant DNA techniques have also made it possible to construct poliovirus type 1/type 3 antigenic hybrid viruses in cell cultures to elicit both type 1 and type 3 neutralizing antibodies. Three poliovirus serotypes are known, and each serotype displays three neutralizing antigenic sites. A type 1/type 3 chimeric virus constructed from clones of Sabin strain derivatives would be useful in the primary vaccination of children to reduce vaccine-associated cases of poliomyelitis. *See* ANIMAL VIRUS; POLIOMYELITIS.

Diarrheal disease vaccines. Diarrheal diseases are caused by the double-stranded, segmented RNA enteroviruses, by the single-stranded RNA Norwalk-type caliciviruses, and by enterotoxigenic bacteria, such as *Vibrio cholerae, Escherichia coli*, and *Shigella dysenteriae*. These pathogens cause about 2.5 million deaths each year in children under 5 years of age, especially in developing countries where food and water are frequently contaminated. Inactivated vaccines and purified toxoids have traditionally been used to prevent cholera (Table 1) and other diarrheal diseases. Recently, however, genetic engineering techniques have enabled scientists to develop new oral vaccines, such as those obtained by genetically engineering the deletion of virulence factors from *Shigella*, *Salmonella*, and *V. cholerae*. *See* CHOLERA; DIARRHEA; ENTEROVIRUS; ESCHERICHIA; GENETIC ENGINEERING; RIBONUCLEIC ACID (RNA).

Rotavirus vaccines. Rotaviruses are the most common cause of severe diarrhea and dehyration in infants. In the developing countries, rotavirus diarrhea causes about 700,000 deaths per year. In the United States, about one-third of pediatric hospital admissions are due to rotaviruses, with about 20 deaths per year. Rotaviruses elaborate an enterotoxin (nonstructural protein NSP4) that triggers a signal transduction pathway that alters epithelial cell permeability and chloride secretion. Natural protective immunity against severe rotavirus disease is built up during the first 2–3 years of life. Studies with live attenuated oral rotavirus vaccines have also shown that the majority of severe episodes of rotavirus diarrhea are preventable by oral immunization. An oral tetravalent rotavirus vaccine was licensed in the United States in August 1998. This vaccine consists of a live-attenuated Rhesus rotavirus serotype 3 combined with human-rhesus reassortants expressing serotypes 1, 2, and 4. The federal government recommended that every Infant in the United States get the protective vaccine. Over 1 million children swallowed three doses of the vaccine, at 2, 4, and 6 months of age. However, in October 1999 the recommendation for routine vaccination of United States infants was withdrawn, and the manufacturer voluntarily withdrew the vaccine from the market, after it was found that 56 children developed a painful and potentially fatal bowel obstruction within 1 week of receiving the vaccine. Several other candidate rotavirus vaccines, including bovine-human reassortants, are currently being investigated. *See* INFANT DIARRHEA.

Oral transgenic plant vaccines. These vaccines exemplify the most recent innovative approach to vaccine development. Ultimately, the vaccine designers hope to use common, uncooked, edible plants, such as bananas and tomatoes, as transgenic plant delivery systems to express the antigens of pathogenic organisms in the gut, so as to induce secretory IgA, the predominant form of immunoglobulin found in mucosal secretions. Presently, however, only model studies have been completed using transgenic potatoes and tobacco plants. Transgenic plants have many potential advantages as vaccines. They are easy to grow at minimal cost, even in the most underdeveloped countries; and unlike bacteria and animal cells, they do not require specialist media and equipment or stringent purification protocols. Transgenic plants can be conveniently self-fertilized to produce stable, true-breeding lines propagated by conventional horticultural techniques and stored or distributed as seeds. Neither sterile syringes nor nurses would be needed for vaccination. *See* BREEDING (PLANT); SOMATIC CELL GENETICS.

Human and animal diarrheas were initially targeted to demonstrate the feasibility of oral transgenic plant vaccines. Transgenic potato and tobacco plants were engineered to express the heat-labile subunit of *E. coli* enterotoxin LT-B (LT stands for heat-labile toxin) and the related subunit of the *V. cholerae* enterotoxin (CT)-B, both of which have affinity for the cell surface receptor, monosialosyl ganglioside GM1. Transgenic plants expressing the capsid protein of the Norwalk calicivirus were also constructed. To optimize the plant cells' expression of the bacterial and viral proteins, a constitutive plant promoter from cauliflower mosaic virus was used to drive transcription, and a plant-specific termination sequence derived from

a soybean vegetative storage protein gene was inserted into the constructs. Fecal IgA secretory antibodies and serum IgG antibodies, were induced by mice fed (via a stomach tube) transgenic plants. Mice immunized with potato LT-B were also partially protected after exposure to intact LT enterotoxin. Furthermore, in preliminary human trials, volunteers who ingested the transgenic potatoes expressing the LT-B subunit developed serum and secretory neutralizing antibodies with minimal adverse side reactions. In other studies, a protective antibody response to foot-and-mouth disease virus was induced in mice following oral or parenteral immunization with alfalfa transgenic plants expressing the foot-and-mouth disease virus structural protein VP1; and neutralizing antibodies were induced by mice immunized with leaf extracts from transgenic plants expressing the glycoprotein S polypeptides of swine-transmissible gastroenteritis coronavirus. Likewise, rabies virus glycoprotein G has been expressed in transgenic tomatoes.

Cancer vaccines. Materials commonly termed cancer vaccines are usually immunotherapeutic preparations targeted to increase the immune response to mutated cellular proteins. They are intended to inhibit the growth of previously diagnosed cancers rather than to prevent the appearance of new, spontaneous malignant tumors. However, in cases in which infectious agents are strongly associated with cancer induction, vaccines capable of preventing cancers have been made available. *See* CANCER (MEDICINE); TUMOR VIRUSES.

Viral hepatitis comprises common infectious diseases of humans. There are more than 350 million carriers of the hepatitis type B virus (HBV) in the world, with about 1 million in the United States. It has been estimated that HBV causes about 1 million deaths worldwide, with at least 5000 deaths in the United States, each year. Carriers of HBV are at risk of developing hepatocellular carcinomas many years after the initial HBV infection. Hepatocellular carcinoma is one of the most common cancers of the world (ranking within the top 10). In some parts of Asia, it may be the first or second most common cancer. A nationwide HBV vaccination program was implemented in Taiwan in 1984. Since the initiation of this universal vaccination program, the incidence of hepatocellular carcinoma in Taiwanese children has halved, and the mortality rate for hepatocellular carcinoma has also decreased. Vaccination programs in Korea and Gambia have also decreased the incidence of hepatocellular carcinoma. *See* HEPATITIS.

Marek's disease virus is a cell-associated herpesvirus of poultry and causes a naturally occurring contagious malignant lymphoma in chickens. The disease has been effectively controlled for more than a decade using a live vaccine based on a non-pathogenic serologically related herpesvirus of turkeys. Cancers caused by other known infectious agents (such as, Epstein-Barr virus, Hepatitis C virus) may also ultimately be preventable by vaccination. *See* AVIAN LEUKOSIS; EPSTEIN-BARR VIRUS. Saul Kit

Bibliography. C. J. Arntzen, Pharmaceutical foodstuffs: Oral immunization with transgenic plants, *Nature Med. Vaccine Suppl.*, 4:502–503, 1998; D. Baxby, *Jenner's Smallpox Vaccine: The Riddle of Vaccinia Virus and Its Origin*, 1981; B. S. Blumberg, Hepatitis B virus, the vaccine, and the control of primary cancer of the liver, *Proc. Nat. Acad. Sci. USA*, 94:7121–7125, 1997; J. C. Cox and A. R. Coulter, Adjuvants: A classification and review of their modes of action, *Vaccine*, 15:248–256, 1997; B. Duff and P. Duff, Hepatitis A vaccine: Ready for prime time, *Obst. Gynecol.*, 91:468–471, 1998; J. L. Melnick, Virus vaccines: Principles and prospects, *Bull. WHO*, 67:105–112, 1989; S. A. Plotkin and E. A. Mortimer, Jr. (eds.), *Vaccines*, 1988; C. O. Tachet et al., Immunogenicity in humans of a recombinant bacterial antigen delivered in a transgenic potato, *Nat. Med.*, 4:607–609, 1998; *Vaccine*, periodical published by Butterworth, London, six times a year.

Vacuole

An intracellular compartment, bounded by a single membrane bilayer, which functions as a primary site of protein and metabolite degradation and recycling in animals, but serves additional complex functions in fungi and plants (**Fig. 1**). Scientists who study vacuoles also define them as the terminal product of the secretory pathway. The secretory pathway functions to transport protein and metabolite-containing membrane vesicles from sites of synthesis or uptake to the vacuole. *See* CELL MEMBRANES; CELL METABOLISM.

Animal lysosomes. In animals, a lytic vacuole known as the lysosome typically functions to process macromolecules. Lysosomes contain an acidic environment and store a large set of specialized hydrolytic enzymes that serve in the breakdown of nucleic acids, sugars, lipids, and proteins. Such macromolecules can be targeted to the lysosome from sites of synthesis. For example, proteins that assemble incorrectly in the endoplasmic reticulum (ER) can be degraded in the lysosome and their

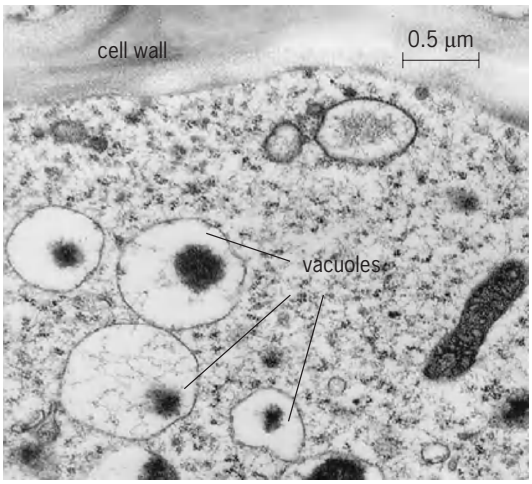


Fig. 1. Electron micrograph of a barley root tip cell, showing multiple vacuoles within the cytoplasm.

constituent amino acids recycled. Proteins that can serve as nutrients are also targeted to the lysosome from the cell surface as a result of endocytosis. An important process for the recycling of cytoplasm in eukaryotic cells is autophagy, in which molecules or organelles are encapsulated in membrane vesicles that fuse with the lysosome. *See* ENDOCYTOSIS; ENDOPLASMIC RETICULUM; LYSOSOME.

In the mammalian immune system, macrophages and neutrophils take up particles and pathogens in the process of phagocytosis, during which the pathogen is eventually digested in the lysosome. A number of diseases in humans can be caused when intracellular pathogens evade destruction in the lysosome. *Mycobacteria* species, which are responsible for severe illnesses such as tuberculosis and leprosy, undergo phagocytosis by cells of the immune system but are able to interfere with the process of transport to the lysosome for digestion. Thus, these bacteria are able to circumvent the immune system and reside as intracellular parasites. *See* PHAGOCYTOSIS.

Fungi. In fungi such as the yeast *Saccharomyces cerevisiae*, vacuoles can serve additional functions not found in animals. Besides a lytic function, they serve in the storage of ions such as sodium, calcium, potassium, and chloride, as well as amino acids for protein synthesis. Ion storage and transport across the vacuolar membrane is particularly important for maintaining the optimal osmotic balance within the cytoplasm for cellular processes. In yeast, it has been demonstrated that vacuoles can also function in the destruction and recycling of cellular organelles, such as peroxisomes, which help protect the cell from toxic oxygen-containing molecules such as peroxides. The process of peroxisome digestion by vacuoles is known as pexophagy. In fungi, many mutants are known to have defective vacuoles or no vacuoles, indicating that vacuoles are not essential under all conditions.

Plants. The most complex vacuoles are found in plants. In contrast to the vacuoles found in yeast and animals, plants contain multiple types of vacuoles with distinct functions essential for survival. Some vacuoles contain hydrolytic enzymes and store ions similar to those found in lysosomes, whereas others serve a role in storing pigments such as anthocyanins which impart color to flowers to attract insect or bird pollinators. Specialized ER-derived vacuoles in plant seeds, known as protein bodies, function in the storage of proteins called prolamines that are common in the endosperm of cereals such as rice and maize. Upon germination, the proteins are degraded and used as a source of amino acids and nitrogen for the growing plant. This is an effective survival strategy, as is the storage in seeds of toxins such as lectins and other macromolecules, such as protease inhibitors, that interfere with predation by herbivores. Such compounds are known as allelochemicals. Other toxins, such as alkaloids, are stored in vacuoles in parts of the plant, such as the leaves, which are subject to frequent herbivory. Scientists have learned that plants produce and store in their vacuoles a vast array of unique chemicals which may, in addition to their natural functions, have medicinal value. *See* PLANT CELL.

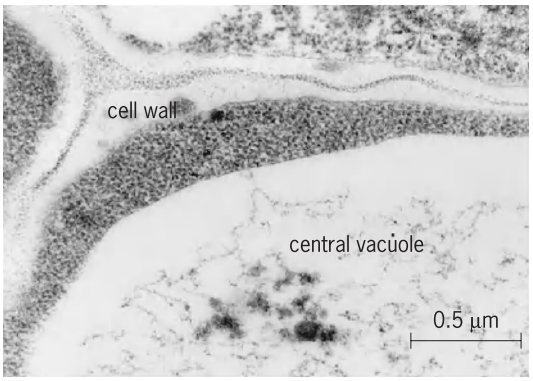


Fig. 2. Electron micrograph of a vacuole from an *Arabidopsis* root tip cell. Such large central vacuoles can accumulate osmotic pressure sufficient to provide turgor for plant cell growth.

Another unique function that vacuoles serve in plants is in cell growth. As a consequence of the accumulation of ions, metabolites, and water, plant vacuoles are under considerable internal osmotic pressure (**Fig. 2**). The vacuolar membrane in plants, known as the tonoplast, as well as the cell itself would burst under this pressure if not for the rigid wall that surrounds the cells. The resulting turgor pressure provides mechanical stability to plant stems. Loss of osmotic pressure in the vacuole due to a lack of water results in plant wilting. The osmotic pressure of the vacuoles also provides the driving force that allows plants to grow by enlarging their cell volume. Enzymes reduce the rigidity of the cell wall, which permits cell expansion under the force of turgor. This is a fundamental process in plants and explains why vacuoles can occupy as much as 95% of the volume of some cells. In fact, the turgor is sufficient to drive rapid movements. In the sensitive plant *Mimosa pudica*, rapid uptake and release of water in vacuoles of specialized cells known as pulvini are responsible for the folding of leaves when touched. *See* CELL WALLS (PLANT).

Transport to the vacuole. Vacuoles are part of a complex secretory pathway that transports protein-containing membrane vesicles from the ER to the plasma membrane or vacuole. In recent years, scientists have made considerable efforts to understand the process by which proteins are targeted to vacuoles. Proteins destined for vacuoles are synthesized at the ER, appropriately processed, then packaged into membrane transport vesicles. The vesicles fuse with the Golgi apparatus, where the proteins are repackaged into transport vesicles that are targeted to the vacuole via another membrane compartment known as the prevacuolar compartment. Each step in this process from ER to Golgi to vacuole requires the formation of new transport vesicles and the subsequent fusion of those vesicles with the next compartment. *See* GOLGI APPARATUS.

The mechanism of vesicle transport is explained by the SNARE hypothesis: Specialized proteins on

transport vesicles known as v-SNARES interact with compatible proteins known as t-SNAREs on target membranes. This interaction leads to membrane fusion. Vesicle formation requires additional specialized proteins. Although many of the basic mechanisms of protein transport are conserved between animals, fungi, and plants, some aspects are quite different. Proteins destined for the vacuole possess signals specifying their transport. In animals the sugar mannose phosphate is frequently used as a signal, whereas in fungi the targeting information resides within the sequence of amino acids specifying the protein. The signal is located at the amino-terminal end of the protein.

Plants have a diverse array of targeting signals, specified by the order of their amino acids. In fact, three different types of signals are found within plant proteins, which are located at the amino terminus, at the carboxyl terminus, or within an internal region of the protein. Other types of macromolecules appear to be transported to the vacuole via specific transport proteins located in the vacuolar membrane. For example, water is moved in and out of the vacuole via transport proteins known as aquaporins. Glenn R. Hicks; Natasha Raikhel

Bibliography. N. J. Bryant and T. H. Stevens, Vacuole biogenesis in *Saccharomyces cerevisiae:* Protein transport pathways to the yeast vacuole, *Microbiol. Mol. Biol. Rev.*, 62(1):230–247, 1998; A. Haas, Reprogramming the phagocytic pathway: Intracellular pathogens and their vacuoles, *Mol. Mem. Biol.*, 15:103–121, 1998; D. J. Klionsky and Y. Ohsumi, Vacuolar import of proteins and organelles from the cytoplasm, *Annu. Rev. Cell Dev. Biol.*, 15:1–32, 1999; S. Kornfeld, The biogenesis of lysosomes, *Annu. Rev. Cell Biol.*, 5:483–525, 1989; F. Marty, Plant vacuoles, *Plant Cell*, 11:587–599, 1999; A. A. Sanderfoot and N. V. Raikhel, The specificity of vesicle trafficking: Coat proteins and SNAREs, *Plant Cell*, 11:629–641, 1999; M. Wink, The plant vacuole: A multifunctional compartment, *J. Exp. Bot.*, 44(suppl.):231–246, 1993.

Vacuum cleaner

An electrically powered mechanical appliance for the dry removal of dust and loose dirt from rugs, fabrics, and surfaces. Portable vacuum cleaners are widely used in domestic and industrial cleaning of surfaces that cannot be wiped or brushed, such as carpets, upholstery, tapestry, or highly contoured surfaces. Small hand vacuum cleaners are used for cleaning the inside of passenger cars. Large industrial tank cleaners are used in cleaning shops. Power saws may be equipped with vacuum cleaners which collect sawdust as it is made. Schools and offices have systems of centrally located dust collectors and fans for creating vacuum; pipes lead to room outlets at which cleaning appliances are attached.

In operation, atmospheric pressure forces air which carries with it loose dirt and dust into the cleaning tool. The dirt-laden air then passes through a bag which retains the dirt. The air is exhausted through an electric fan and an additional filter. The fan provides the low internal air pressure, commonly called the vacuum.

Design of this machine centers on convenience of handling, features including a flexible air hose and tools for cleaning special surfaces and crevices, with considerable variety in location or motion of various assemblies. A critical feature is the filter action of the bag, which may be of cloth (in which case it can be emptied) or of paper (in which case it is disposable). Because at the exhaust the fan provides air at a slight pressure, the appliance may include attachments for spraying.

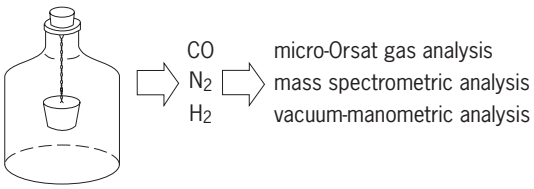
The great advantage of the vacuum cleaner is its retention of dirt in the filter bag. However, the force of the draft is limited by the use of atmospheric pressure. For this reason, local industrial cleaning of work in process may be by compressed air. *See* COMPRESSOR; VACUUM PUMP. Frank H. Rockett

Vacuum fusion

A technique of analytical chemistry for determining the oxygen, hydrogen, and sometimes nitrogen content of metals. The method can be applied to a wide variety of metals, the alkali and alkaline earth metals being exceptions. The range of the method extends from 1% down to a few parts per million for oxygen and nitrogen and down to fractional parts per million for hydrogen.

The metal sample is either fused or dissolved in a bath, or flux, of a second metal in a heated graphite crucible supported inside an evacuated glass or quartz vessel (see **illus.**). Oxygen is released from the metal as carbon monoxide by reaction of oxides or dissolved oxygen with carbon from the graphite crucible at high temperature. Metal nitrides dissociate, although not always quantitatively, to form elemental nitrogen. Hydrogen is evolved as elemental hydrogen.

The mixture of carbon monoxide (CO), nitrogen (N_2), and hydrogen (H_2) is analyzed to determine individual component concentrations by one of several techniques, including micro-Orsat, mass-spectrometric, and vacuum-manometric procedures. The last is in most general use. Gas quantities are determined by measurement of the pressure of the gas after confinement into a small calibrated volume. The product of the pressure and volume is



Vacuum-fusion methods of analysis.

proportional to the total number of moles of gas present, regardless of the species of gas molecules. Measurement is first made of the total quantity of CO, N_2, and H_2 collected. These gases are then oxidized to carbon dioxide (CO_2), water (H_2O), and N_2 by passing over hot copper oxide. The CO_2 and H_2O can be removed from the gas mixture successively by selective freezing or by use of chemical absorbents. The decrease in total moles of gas accompanying removal of each species permits determination of the number of moles of CO_2, N_2, and H_2O.

Furnace temperatures and bath conditions must be selected for each metal or alloy to ensure quantitative recovery of oxygen. The required furnace temperatures range from 1650°C (3000°F) for iron, nickel, and low-alloy steels to as high as 1900°C (3500°F) for metals such as titanium, zirconium, tantalum, and thorium. Iron, nickel, and low-alloy steels are usually fused without addition of other metals. High-melting metals require use of a second metal to lower their melting points. The second metal is used as a previously prepared bath or, in some instances, is added simultaneously with the sample metal. Iron, nickel, platinum, and tin are the most generally used fluxing metals.

Nitrogen is not always quantitatively evolved, and values obtained by vacuum-fusion methods for nitrogen are not generally reliable. Hydrogen is always evolved quantitatively upon fusion of the metal in a vacuum. When hydrogen alone is of interest, simpler and more rapid techniques can be used, involving vacuum extraction from metal heated to temperatures well below the melting point.

Vacuum-fusion analysis of gases and metals has found its widest industrial application, from a process-control viewpoint, in titanium, tantalum, niobium, zirconium, and molybdenum. Determination of hydrogen in steel is a routine procedure in some applications. As a research tool, the vacuum-fusion techniques have been applied to practically all of these materials except the alkali and alkaline-earth metals.

Another method, inert gas fusion, was developed for determination of gases in metals. The techniques used are similar to vacuum fusion but substitute inert gases for the vacuum environment. This method is sometimes referred to as argon-fusion. The methods of detection of the gases involved during the fusion are modified somewhat. *See* GAS AND ATMOSPHERE ANALYSIS; VACUUM METALLURGY. Frank C. Benner

Bibliography. A. Choudhury, *Vacuum Metallurgy*, 1990; J. D. Winefordner and M. M. Bursey (eds.), *Treatise on Analytical Chemistry*, vol. 11, pt. 1: *Theory and Practice*, 2d ed., 1989.

Vacuum measurement

The determination of a gas pressure that is less in magnitude than the pressure of the atmosphere. This low pressure can be expressed in terms of the height in millimeters of a column of mercury which the given pressure (vacuum) will support, referenced to zero pressure. The height of the column of mercury which the pressure will support may also be expressed in micrometers. The unit most commonly used is the torr, equal to 1 mm (0.03937 in.) of mercury (mmHg). Less common units of measurement are fractions of an atmosphere and direct measure of force per unit area. The unit of pressure in the International System (SI) is the pascal (Pa), equal to 1 newton per square meter (1 torr = 133.322 Pa).

Atmospheric pressure is sometimes used as a reference, particularly for relatively crude measurements. The pressure of the standard atmosphere is 29.92 in. or 760 mm of mercury (101,325 Pa or 14.696 lbf/in.^{2}); therefore, using atmospheric pressure as a reference, an absolute pressure of 10 in. of mercury would be expressed as 19.9 in. of mercury vacuum.

In the laboratory, measurement of vacuum is important because the vacuum level has a significant effect on most physical, chemical, and biological processes.

In industry, vacuum level is commonly measured and controlled to maintain uniformity of product and as a guide to safe operation. Vacuum measurement is used, for example, in the manufacturing of television sets (evacuation of picture tubes); in metallurgy (treatment of metals attacked by common gases); and in the pharmaceutical industry (distilling heat-sensitive compounds). *See* PROCESS CONTROL.

Pressures above 1 torr can be easily measured by familiar pressure gages, such as liquid-column gages, diaphragm-pressure gages, bellows gages, and bourdon-spring gages. At pressures below 1 torr, mechanical effects such as hysteresis, ambient errors, and vibration make these gages impractical. *See* MANOMETER; PRESSURE MEASUREMENT.

Pressures below 1 torr are best measured by gages which infer the pressure from the measurement of some other property of the gas, such as thermal conductivity or ionization. The thermocouple gage, in combination with a hot- or cold-cathode gage (ionization type), is the most widely used method of vacuum measurement today. *See* IONIZATION GAGE.

Other gages used to measure vacuum in the range of 1 torr or below are the McLeod gage, the Pirani gage, and the Knudsen gage. The McLeod gage is used as an absolute standard of vacuum measurement in the $10–10^{-4}$ torr ($10^3–10^{-2}$ Pa) range. *See* MCLEOD GAGE; PIRANI GAGE.

The Knudsen gage is used to measure very low pressures. It Measure pressure in terms of the net rate of transfer of momentum (force) by molecules between two surfaces maintained at different temperatures (cold and hot plates) and separated by a distance smaller than the mean free path of the gas molecules. Richard Comeau

Bibliography. T. A. Delchar, *Vacuum Physics and Technology*, 1993; D. Hucknall, *Vacuum Technology and Applications*, 1993.

Vacuum metallurgy

The making, shaping, and treating of metals, alloys, and intermetallic and refractory metal compounds in a gaseous environment where the composition and partial pressures of the various components of the gas phase are carefully controlled. In many instances, this environment is a vacuum ranging from subatmospheric to ultrahigh vacuum (less than 760 torr or 101 kilopascals to 10^{-12} torr or 10^{-10} pascal). In other cases, reactive gases are deliberately added to the environment to produce the desired reactions, such as in reactive evaporation and sputtering processes and chemical vapor deposition. The processes in vacuum metallurgy involve liquid/solid, vapor/solid, and vapor/liquid/solid transitions. In addition, they include testing of metals in controlled environments. The **table** lists the processes, pressure ranges, and gas environments.

The early applications of vacuum metallurgy were involved with the more conventional aspects, such as extractive metallurgy, melting, and heat treatment. Since the 1970s, two important developments have occurred. First, other nonvacuum specialty melting processes, such as electroslag remelting (ESR) and plasma melting, have been developed. They have some unique aspects (for example, grain size refinement) which make them strong competitors with vacuum melting processes on the one hand and complementary melting processes on the other. For many applications, the final ingot may have been sequentially processed by a number of vacuum and nonvacuum melting processes. Second, in many high-technology applications, a single material, such as a metal or alloy, cannot meet the multiple requirements placed on it for a particular application. For example, the blades and vanes for the hot stage of a gas turbine engine have to be a composite material, with high-temperature nickel or cobalt base alloy for strength at elevated temperatures and a coating of an alloy such as nickel (or cobalt)-chromium-aluminum-yttrium to resist the severe corrosion environment. Thus a large and rapidly growing aspect of vacuum metallurgy consists of vapor deposition techniques.

Advantages. There are three basic reasons for vacuum processing of metals: elimination of contamination from the processing environment, reduction of the level of impurities in the product, and deposition with a minimum of impurities. Contamination from the processing environment includes the container for the metal and the gas phase surrounding the metal. For example, melting of titanium in a refractory crucible in air would severely contaminate the melt both from the material of the crucible and from the atmospheric oxygen, making it useless. Therefore titanium is processed in water-cooled copper crucibles (which do not react with the metal) and in vacuum. In the vacuum process, impurities, particularly oxygen, nitrogen, hydrogen, and carbon, are released from the molten metal and pumped away; and metals, alloys, and compounds are deposited with a minimum of entrained impurities.

There are numerous and varied application areas for vacuum metallurgy; some of these are discussed below.

Various metallurgical processes and their operating pressure regimes*

Processes	Approximate pressure regime, torr (Pa)	Gas environment†
Extraction and refining of metals	Several atm to 10^{-3} torr (0.133)	V
Degassing of liquid steel	400 to 1 (53,200 to 133)	V
Vacuum induction melting	100 to 10^{-6} (13,300 to 1.33×10^{-4})	V
Vacuum arc melting	1 to 10^{-3} (133 to 0.133)	V, I
Electron-beam melting	10^{-3} to 10^{-8} (0.133 to 1.33×10^{-6})	V
Levitation melting	760 to 10^{-5} (101,080 to 1.33×10^{-3})	V, I
Melting on water-cooled hearths	760 to 10^{-5} (101,080 to 1.33×10^{-3})	V, I
Zone refining	760 to 10^{-8} (101,080 to 1.33×10^{-6})	V, I
Surface treatment (carburizing, nitriding)	10^{-1} to 10^{-3} (13.3 to 0.133)	R
Heat treatment:		
High vacuum	1 to 10^{-6} (133 to 1.33×10^{-4})	V
Ultrahigh vacuum	10^{-8} to 10^{-10} (1.33×10^{-6} to 1.33×10^{-8})	V
Evaporation		
For purification and bulk deposits	10^{-2} to 10^{-7} (1.33 to 1.33×10^{-5})	V
For thin-film deposition	10^{-1} to 10^{-10} (13.3 to 1.33×10^{-8})	V, R
Sputtering to form thin-film or bulk deposits	10^{-2} to 10^{-5} (1.33 to 1.33×10^{-3})	I, R
Chemical vapor deposition	760 to 10^{-2} (101,080 to 1.33)	V, R
Electron-beam welding (except where beam is brought out to atmosphere)	$<10^{-4}$ ($<1.33 \times 10^{-2}$)	V
Brazing	10^{-2} to 10^{-5} (1.33 to 1.33×10^{-3})	V
Diffusion bonding	10^{-2} to 10^{-5} (1.33 to 1.33×10^{-3})	V
Evaluation of physical and mechanical properties	Several atm to 10^{-9} torr (1.33×10^{-7})	V
Study of nature of surfaces and reactions occurring on clean surface	$<10^{-8}$ ($<1.33 \times 10^{-6}$)	V
Mechanical working in vacuum	10^{-1} to 10^{-4} (13.3 to 1.33×10^{-2})	V
Sintering of powders in vacuum	10^{-1} to 10^{-5} (13.3 to 1.33×10^{-3})	V
Electron-beam machining	$<10^{-4}$ ($<1.33 \times 10^{-2}$)	V

*After R. F. Bunshah, *Techniques of Metals Research*, vol. 1, John Wiley and Sons, 1968.
†V = residual gases in a vacuum; I = inert gas added; R = reactive gas added.

Extractive metallurgy. Included in extractive metallurgy are the beneficiation of ores, reduction of ores or compounds to crude metals, and subsequent refining of the high-purity metals. Examples are the reduction of dolomite with ferro-silicon, refining of rare-earth metals by distillation, and preparation of pure metals by halide dissociation on a heated wire.

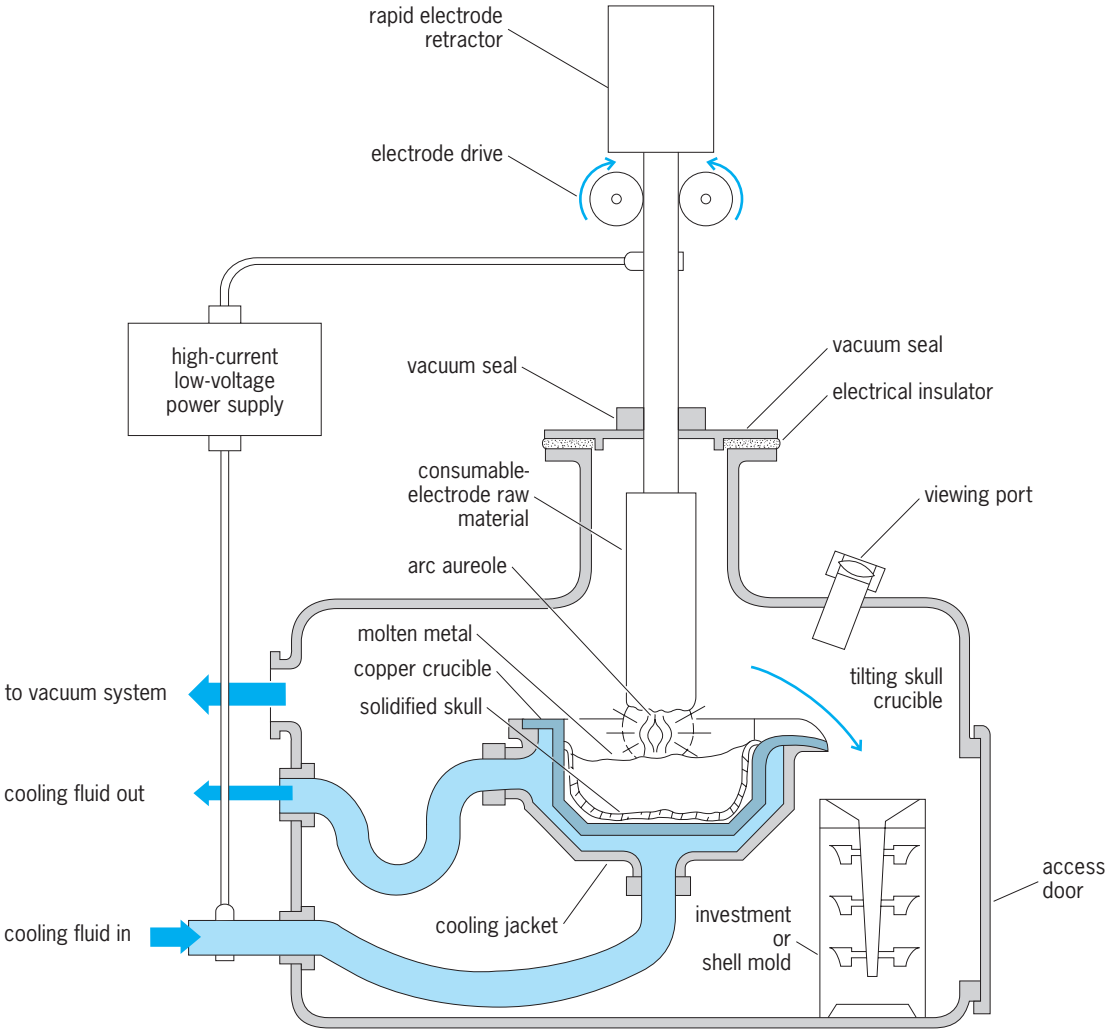
Melting processes. The melting processes have several functions. Among them are the refining of crude metals, the consolidation of various components into alloys, and the production of an ingot of desired shape.

The metal may be heated and melted by several heating methods which distinguish the various processes, for example, induction melting, consumable-electrode arc melting (see **illus.**), and electron-beam melting. In addition, there is the levitation melting technique where no crucible is used, the metal being supported by an electromagnetic force.

Vacuum melting and allied processes are (1) titanium alloy scrap recycling to gravimetrically separate high-density inclusions such as broken tungsten carbide tool bits and other heavy impurities; (2) removal of low-density inclusions and production of a titanium alloy ingot from scrap turnings, scrap solids, virgin sponge, aluminum, and master alloys; (3) development of a single merit process, vacuum-electroslag, to replace the two melt process for high-temperature superalloys; it uses a thin slag layer of calcium oxide/aluminum oxide on the molten pool, which reduces nitrogen and hydrogen pick-up and oxidation of the metal from the slag; melt cleanliness is improved since the slag acts as a filter; and (4) nitrogen-alloyed steels as a substitute for the more expensive nickel-alloyed steels.

Casting of shaped products. The molten metal is cast into intricate shapes by using a mold of refractory compounds usually made by the lost-wax process. A good example of this process is the production of components for aircraft jet engines by using a nickel-base high-temperature alloy. *See* METAL CASTING.

Vacuum degassing of molten steel. This is the largest single application of vacuum metallurgy. In this process, steel melted in conventional electric arc



Vacuum metallurgy. Skull-crucible, consumable-electrode vacuum arc melting and casting furnace. (*After R. F. Bunshah, Techniques of Metals Research, vol. 1, John Wiley and Sons, 1968*)

furnaces is treated in vacuum before being cast into ingots. There are several processes for vacuum degassing. During the process the gases in the molten metal are released and pumped away. *See* STEEL; STEEL MANUFACTURE.

Heat treatment. Heat treatment of refractory and reactive metals (titanium, columbium, tungsten, and so forth) is carried out in vacuum to promote desired reactions in the solid metal while preventing recontamination from the atmosphere; an example is age-hardening of titanium alloys to get maximum strength.

Surface treatment. Modification of the surface properties of a material is carried out by diffusing various species such as carbon, nitrogen, boron, or aluminum from a solid, liquid, or gaseous environment. The effects of impurities in the gas phase will affect the surface properties, and hence the need for a controlled environment. In some instances ionization of the reactive gas permits the process to occur more rapidly and at lower temperatures, for example, ion nitriding and ion carburizing.

Joining. Joining processes include welding, brazing, and diffusion bonding. They are carried out in vacuum where the metals being joined would be contaminated by joining in air. A development of particular importance is electron-beam welding. *See* ARC WELDING; WELDING AND CUTTING OF MATERIALS.

Vapor deposition. This type of process includes the physical vapor deposition—evaporation, ion plating, and sputtering—and the chemical vapor deposition. The products of these processes are coatings and free-standing shapes such as sheet, foil, and tubing of thickness ranging from 20 nanometers to 25 mm. These processes have widespread application in diverse applications in the metallurgical, chemical, engineering, and electronic industries. A large variety of products are made possible by using high-vacuum coating processes.

Decorative. These include automotive trim (interior and exterior), toys, cosmetic packaging, pens and pencils, Christmas decorations, food and drink labels, costume jewelry, home hardware, eyeglass frames, packaging and wrapping materials, and watch cases.

Optically functional. These include laser optics (reflective and transmitting), architectural glazing, home mirrors, automotive rearview mirrors, eyeglass lenses, projector reflectors, camera lenses and filters, instrument optics, auto headlight reflectors, television camera optical elements, and meter faces.

Electrically functional. Included in this group are semiconductor devices, integrated circuits, capacitors, resistors, magnetic tape, disk memories, superconductors, electrostatic shielding, switch contacts, and solar cells.

Mechanically functional. These products are aircraft engine parts, aircraft landing gear, solid film lubricants, and tool bit hard coatings.

Chemically functional. In this group are corrosion-resistant fasteners, gas turbine engine blades and vanes, battery strips, and marine equipment.

Ultrafine powders. Ultrafine powders of metals and ceramics, 1–10 nm in diameter, are produced by converting the condensed metal phase to a vapor by evaporation or sputtering processes at inert or inert-plus-reactive gas pressures at about 100 millitorr so as to produce gas-phase condensation into ultrafine particles, which are then trapped on liquid-nitrogen-cooled surfaces. If a ceramic compound (oxide, carbide, nitride) powder is desired, the gas phase contains the appropriate reactive base (oxygen, hydrocarbon, or nitrogen) to form the compound.

Space processing. The space environment combines microgravity conditions and high vacuum. Results of experimental work in various space missions have demonstrated improvements in properties of metals, semiconductors, and glasses over similar material processed on Earth in a 1-*g* environment. The space environment also presents a more functional possibility of studying physical and chemical properties of materials. In a particular configuration known as the molecular shield, an extreme vacuum environment (less than 10^{-13} torr) with a very high pumping throughput is available in space. Such an environment is not possible on Earth. When combined with the micro-*g* environment of space, the potential for producing entirely unique materials is large. *See* SPACE PROCESSING.

Other processes. Other metallurgical operations carried out in vacuum are the testing of metals, electron-beam machining, study of surfaces, powder metallurgy, and mechanical working. *See* METAL COATINGS; POWDER METALLURGY.

The products of vacuum processing include the entire gamut of materials, steels, reactive metals (titanium, zirconium, and so forth), refractory metals (tungsten, molybdenum, and others), their alloys, intermetallic compounds, and refractory compounds. Vacuum metallurgy occupies a key role in the design of space processing programs. *See* PYROMETALLURGY; PYROMETALLURGY, NONFERROUS.

Rointan F. Bunshah

Bibliography. American Vacuum Society, *Transactions of the Vacuum Metallurgy Conference*, 1967, also other annual editions; J. A. Belk, *Vacuum Techniques in Metallurgy*, 1963; R. F. Bunshah, *Deposition Technologies for Films and Coatings: Developments and Applications,* 2d ed., 1994; R. F. Bunshah, History and current status of vacuum metallurgy, *J. Vac. Sci. Technol.*, A12(4):936–945, 1994; R. F. Bunshah (ed.), *Vacuum Metallurgy*, 1958; *J. Vac. Sci. Technol.,* 1964 to date.

Vacuum pump

A device that reduces the pressure of a gas (usually air) in a container. When gas in a closed container is lowered from atmospheric pressure, the operation constitutes an increase in vacuum in this container. The unit used for vacuum is a millimeter (mm) column of mercury (Hg). Another term for the mmHg unit is the torr. These terms are used interchangeably. The torr is named in honor of E. Torricelli, a

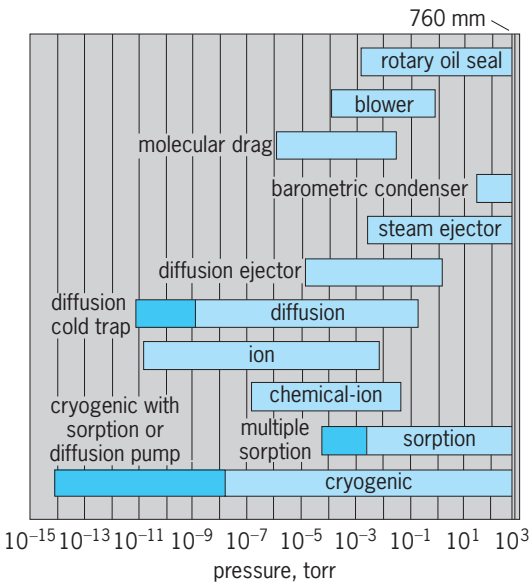


Fig. 1. Comparison of the degree of vacuum attained by important types of vacuum pumps.

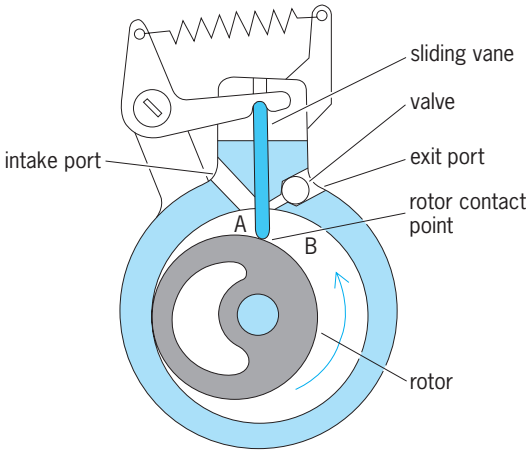


Fig. 2. Chief components of rotary oil-seal pump.

pioneer in vacuum technology. The various vacuum regions are classified as follows:

Low vacuum, 760 (atmospheric pressure) to 25 mmHg
Medium vacuum, 25 to 10^{-3} mmHg
High vacuum, 10^{-3} to 10^{-6} mmHg
Very high vacuum, 10^{-6} to 10^{-9} mmHg
Ultrahigh vacuum, 10^{-10} mmHg and beyond

A vacuum as high as 10^{-15} mmHg or one-billionth of a billionth of an atmosphere has been reached. *See* PRESSURE.

Vacuum pumps are evaluated for the degree of vacuum they can attain and for how much gas they can pump in a unit of time. **Figure 1** compares the degree of vacuum that the more important types of vacuum pumps can attain. The determination of the quantity of gas a vacuum pump handles is discussed later. At low vacuum, this quantity is usually expressed as pounds of air per hour. Any other gas is converted, via molecular weight relationship, to air equivalent for this purpose. For high vacuums, pounds of air per hour is expressed as torr liters (or mmHg liters) per second. In practice, where high vacuum is required, two or more different types of pumps are used in series. For example, a rotary oil-seal pump or steam ejector can be used with a diffusion pump to have continuity in pumping from atmospheric pressure to the high or very high vacuum range.

Mechanical pumps. One of the early vacuum pumps was similar in design to the reciprocating steam engine. A power-driven piston working in a cylinder sucked the gas out of any attached container. A more recent development is the rotary oil-seal pump (**Fig. 2**). Gas is sucked into chamber A through the opening intake port by the rotor. A sliding vane partitions chamber A from chamber B. The compressed gas that has been moved from position A to position B is pushed out of the exit port through the valve, which prevents the gas from flowing back. The valve and the rotor contact point are oil-sealed. Since each revolution sweeps out a fixed volume, it is called a constant-displacement pump.

A rotary blower pump operates by the propelling action of one or more rapidly rotating lobelike vanes. It does not use oil as a sealing medium.

The molecular drag pump is one that operates at very high speeds, as much as 16,000 revolutions/min. Pumping is accomplished by imparting a high momentum to the gas molecules by the impingement of the rapidly rotating body.

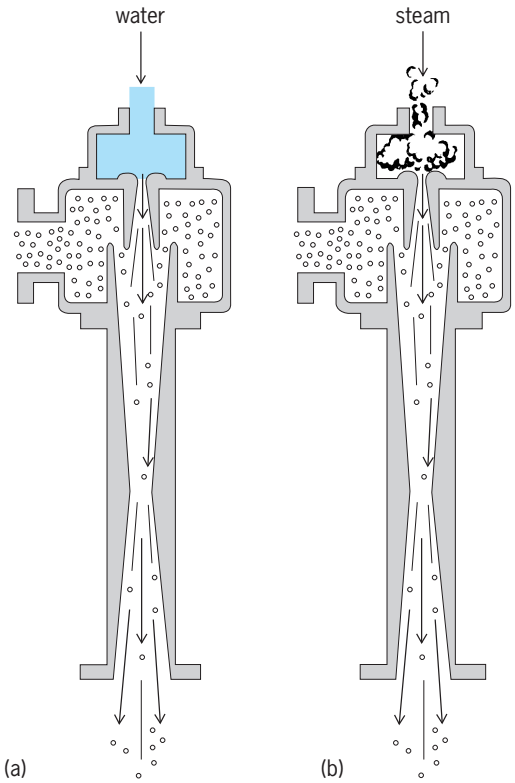


Fig. 3. Diagram of the water aspirator pump. Utilization of (*a*) water and (*b*) steam.

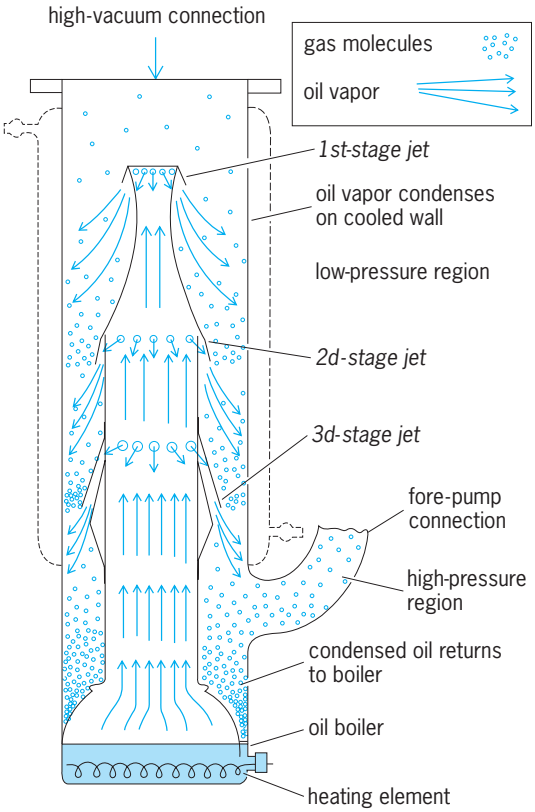


Fig. 4. Main operating features of diffusion pump.

Ejector pumps. The earliest form of this type is the water aspirator (**Fig. 3**). When water is forced under pressure through the jet nozzle, it will force the gas in the inlet chamber to go through the diffuser, thus lowering the pressure in the inlet chamber. When high-pressure steam is used instead of water, it is called a steam ejector. Steam ejectors often are combined in multiple series with barometric condensers to reach lower pressures. A barometric condenser operates by injecting water into the upper end of a vertical 35-ft (11-m) pipe whose exit is immersed in a catch basin called a hot well. The gas trapped in the inlet chamber is carried out with the water. Like the water aspirator, it can produce a vacuum (usually between 25 and 40 mmHg) equal to the vapor pressure of the water. This type of pumping is widely used in the chemical and petroleum industries.

Diffusion pumps permit high pumping speed at very high vacuum. The vapor-jet principle of the ejector pumps has been modified by using oil or mercury vapor diffusing through a jet, then condensing the vapor on the cooled wall of the pump chamber. **Figure 4** illustrates the diffusion pump operation. **Figure 5** shows a small, 1-in.-diameter (2.5-cm) diffusion pump with pumping speeds as low as 0.26 gallon/s (1 liter/s) compared for size with one of the largest pumps, 48 in. (120 cm) in diameter having a pumping speed of 26,000 gallons/s (100,000 liters/s). Sometimes to meet operating requirements one jet is built on the ejector principle. Then it is called a diffusion-ejector pump.

Specialty pumps. A number of pumps have been developed which meet special pumping requirements. In all cases it is not the principles that are new, but the way they are used. The ion pump operates electronically. Electrons that are generated by a high voltage applied to an anode and a cathode are spiraled into a long orbit by a high-intensity magnetic field (**Fig. 6**). These electrons colliding with gas molecules ionize the molecules, imparting a positive charge to them. These are attracted to, and are collected on, the cathode. Thus a pumping action takes place. Ion pumps are used for evacuating nuclear accelerators, x-ray tubes, and other specialized equipment. In a chemical-ion pump, a metal (usually titanium) is evaporated by electrical power, and the metal vapor reacts with chemically active molecules (O_2 and N_2). The reaction products condense on the walls of the pump. The inert gases that remain are removed by devices similar to that of the ion pump. Sorption pumping is the removal of gases by adsorbing and absorbing them on a granular sorbent material such as a molecular sieve held in a metal container. When this sorbent-filled container

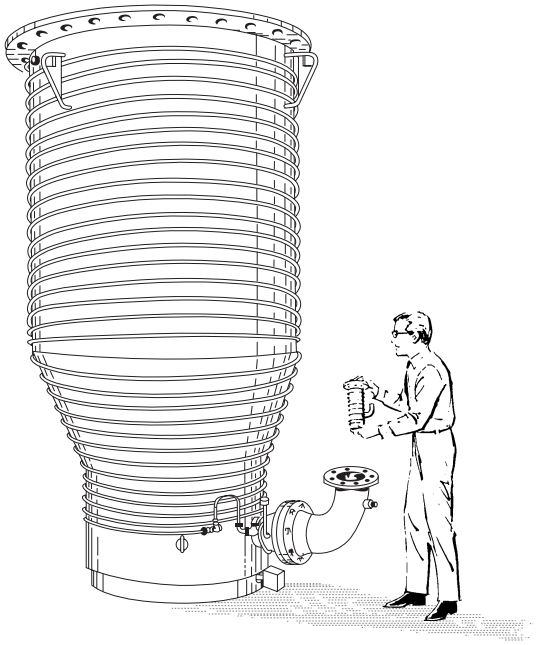


Fig. 5. Size comparison of two diffusion pumps.

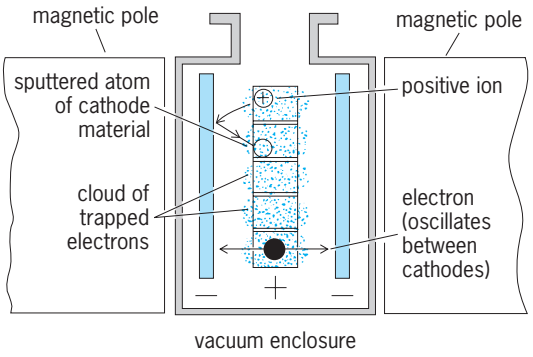


Fig. 6. Principal features of an ion pump.

is immersed in liquid nitrogen, the gas is "sorbed." This type of pump produces a quick, clean vacuum. Cryogenic pumping is accomplished by condensing gases on surfaces that are at extremely low temperatures. This is usually attained by the use of liquid or gaseous hydrogen at around −423°F (−253°C) or liquid helium at −452°F (−269°C). The equipment is very complex in design. However, it attains very high pumping speeds for low-density wind tunnels and space simulator chambers.

Vacuum pumping equations. There are two relations of practical use in vacuum technology. The first is $P_1V_1 = P_2V_2$ (at constant temperature). That is, when pressure P_1 is changed to a new pressure P_2, the volume of the gas at P_1 which is V_1 is changed to a new volume V_2.

The second is $Q = SP$. Here Q is the gas throughput (expressed as torr liters/second at a specified temperature), S is the speed of pumping (in this case liters/second), and P is the pressure (in torr) at the point where S is to be measured. The measurement of S requires a vacuum gage to obtain P and one of several methods for measuring Q. One way to determine S is to measure the amount of air at atmospheric pressure that is admitted into the vacuum chamber through a controlled, calibrated leak. This measurement gives Q, and S is then obtained from the second equation. A more precise method, especially useful for very high and ultrahigh vacuum, is the measurement of a pressure drop across an orifice of known gas conductance. The gas conductance is a mathematically determined quantity obtained from the geometry of the test portion of the vacuum chamber. *See* PUMP.

Edward S. Barnitz

Bibliography. L. G. Carpenter (ed.), *Vacuum Technology: An Introduction*, 2d ed., 1983; A. Roth, *Vacuum Technology*, 3d ed., 1990, reprint 1998.

Vacuum tube

An electron device in which use is made of the electrostatically or magnetically controlled flow of electrons in an evacuated space, or of the phenomena accompanying the electron emission, acceleration, and collection associated with the flow. In common usage, completely evacuated tubes are called hard tubes; tubes containing a small amount of deliberately added gas or vapor are called soft tubes, but they are usually included in the family of vacuum tubes.

Vacuum tubes are used as active and passive electrical circuit elements. In suitable circuitry they rectify, detect, gate, amplify, and oscillate. They can also scan and display images; generate heat, light, and radiation; ionize and accelerate atoms and molecules; and measure high vacuum.

Principle of operation. Most vacuum tubes employ the Edison effect, in which a unidirectional flow of current takes place through a vacuum by thermionic emission of electrons from a hot surface (the cathode). The electrons are collected by a nearby positively charged electrode (the anode or plate). Two-electrode Edison-effect tubes were first employed as radio-frequency detectors, but the enormous potential of the device was not realized until it was shown that the current could be controlled with large amplification ratios and great sensitivity by small electrical charges on a third, meshlike electrode (the grid) interposed between cathode and anode. From the resulting three-element tube (the triode) came all the characteristic inventions of the modern electronic age. *See* ELECTRON EMISSION; ELECTRON MOTION IN VACUUM; THERMIONIC EMISSION.

Later refinements included additional grids by which certain problems inherent in triodes were overcome, and combination tubes were developed in which several functions were combined within one vacuum envelope. Elaborations of the basic idea include camera and picture tubes and magnetrons, klystrons, and traveling-wave tubes for microwave radar. *See* KLYSTRON; MAGNETRON; MICROWAVE TUBE; PICTURE TUBE; RADAR; TELEVISION CAMERA TUBE; TRAVELING-WAVE TUBE.

Use. Originally, vacuum tubes were enclosed in cylindrical glass structures (hence the name tube). Through the walls or base of the tube were sealed the metal electrodes which emitted, controlled, or collected the electrons traversing the empty space within. Up through the 1960s, vast numbers of small tubes were used in radio and television broadcast receivers. They also found their way into industrial instruments and controls, and into the early electronic computers, but their size, fragility, inordinate power consumption, and short lifetimes made them unpopular for these more demanding purposes. Since that time, such vacuum tubes have been almost entirely replaced by the much smaller, cheaper, and longer-lived solid-state devices. Only a few standard receiving-tube types still remain in production, and are used where their electrical characteristics are particularly apt, or where environmental or circuit conditions would destroy solid-state devices. A significant exception is the cathode-ray tube (CRT), which is still made in great numbers for consumer applications. Though bulky, complex, and fragile, it is highly versatile, and so remains in favor as a display device for television, computers, and instruments of all kinds. *See* CATHODE-RAY TUBE; ELECTRONIC DISPLAY.

Other vacuum tubes in widespread use are high-power transmitting and microwave tubes; x-ray and accelerator tubes; mass spectrometer, electron microscope, night-vision, and surface-analytical tubes; and electron beam guns (partial tubes) for evaporators and welders. Since there is no prospect that any of these types will be replaced by solid-state devices, research and development on them continues, and new types occasionally appear. Such specialized tubes range in size from ultraminiature, flat devices sandwiched between glass or sapphire plates for flat displays or high-temperature uses, to giant accelerator chambers used for scientific research. In power-handling ability, they range from microwatts to megawatts. In the larger sizes especially, they may be welded, brazed, or bolted together from

ceramic-metal subassemblies. When fitted with auxiliary vacuum pumps and water-cooling systems, and integrated closely with other circuit elements for use in space, undersea, or military environments, they may be scarcely recognizable as tubes. *See* ELECTRON MICROSCOPE; IMAGE TUBE (ASTRONOMY); ION SOURCES; LIGHT AMPLIFIER; MASS SPECTROSCOPE; PARTICLE ACCELERATOR; PHOTOMULTIPLIER; PHOTOTUBE; SURFACE PHYSICS; WELDING AND CUTTING OF MATERIALS; X-RAY TUBE. Nicholas Reinhardt

Bibliography. J. W. Gewartowski and H. A. Watson, *Introduction to Electron Tubes*, 1969; A. S. Gilmour, Jr., *Microwave Tubes*, 1986; S. Okamura (ed.), *The History of Electron Tubes*, 1994; F. Rosebury, *Handbook of Electron Tube and Vacuum Techniques*, 1965, reprint 1993.

Vaginal disorders

Inflammations, trauma, tumors, congenital defects, and functional disorders of the vagina, the muscular membranous canal from the vulvar opening to the uterine cervix.

Vaginitis. The most common vaginal disorder is vaginitis, an inflammation of varying degree, characterized by itching (pruritus), typically accompanied by a discharge. Breast-feeding infants sometimes display a mild vaginal discharge due to hormones in breast milk. Young children with a vaginal discharge must be examined to exclude the presence of a foreign object or an infectious condition such as gonorrheal vulvovaginitis.

Among women of reproductive age, the three most common types of vaginitis are caused by the protozoon *Trichomonas*, the bacterium *Gardnerella*, or the monilial fungus. Postmenopausal women with little or no estrogen are most likely affected by senile vaginitis due to irritation of the thinned, atrophic mucosa. The diagnosis can usually be made by microscopic inspection of the discharge, and appropriate medication prescribed. Women with new or multiple sexual partners should also have vaginal cultures taken for chlamydia and gonorrhea. Recurrent monilia infections could be a clue to diabetes or human immunodeficiency virus (HIV) infection. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); GONORRHEA.

Other disorders. Trauma of the vagina is common. Lacerations suggest either forced intercourse or the forceful use of foreign objects. Accidental injury as a result of falling on penetrating objects may also occur. Irritation, sensitivity, or even ulceration can occur from ingredients in douches, suppositories, or lubricants, or even from a forgotten tampon.

The most important primary lesion of the vagina is carcinoma, but its incidence is low. Cancer of the cervix is much more common, and in advanced cases the cancer may extend into the vagina. Hyperpigmented areas in the vagina and labia must be monitored carefully to exclude melanoma. Benign lesions such as papillomas, fibromas, and hemangiomas are relatively uncommon and of minor consequence. *See* CANCER (MEDICINE).

Congenital malformations of the vagina are uncommon. They include a double vagina, a transverse septum, and an imperforate hymen. These conditions typically require surgical correction. Congenital absence of the vagina is important as menstrual blood cannot escape and reflexes into the pelvic cavity, a condition leading to the formation of endometriosis. Therefore, a vaginal canal must be established at the time of menarche. Congenital cysts are common but rarely significant.

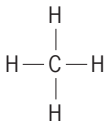
The final category of vaginal disorder is vaginismus, which is characterized by spasmodic, painful, and involuntary contraction of the lower vaginal muscles. Although typically psychological in origin, it can be due to physical causes. Rarely, vaginismus is interpreted as infertility because of the inability to have intercourse. Sexual counseling is often therapeutic, and infertility can be treated by using artificial insemination. *See* REPRODUCTIVE SYSTEM DISORDERS. Machelle M. Seibel

Valence

The combining power of a chemical element for other elements as measured by the number of bonds to other atoms that one atom of the given element forms upon chemical combination; also known as valency. Valence theory concerns all the physical and chemical properties of molecules that especially depend on molecular electronic structure.

Thus, in water, H_2O or

the valence of each hydrogen atom is 1; the valence of oxygen, 2. In methane, CH_4 or



the valence of hydrogen again is 1; of carbon 4. In sodium chloride, NaCl, and carbon tetrachloride, CCl_4, the valence of chlorine is 1, and CH_2 the valence of carbon is 2.

Much more is known about a water molecule than that it contains two hydrogen atoms and one oxygen atom. Each OH distance is 9.57 nanometers and the HOH bond angle is $104^\circ 27'$. The oxygen and hydrogen ends of the molecule are negatively and positively charged, giving it a dipole moment 1.84×10^{-18} electrostatic unit (esu). The molecule absorbs infrared light strongly but is transparent to visible light. Scientists have provided quantitative understanding of these properties and many more in terms of the fundamental theory of quantum mechanics. *See* BOND ANGLE AND DISTANCE; QUANTUM CHEMISTRY.

Combining power of an element. By the 1920s the most important facts about atoms had been established experimentally. A neutral atom of atomic number Z comprises a massive nucleus of charge $+Ze$ and Z very light electrons, each of charge $-e$, where $e = 4.80 \times 10^{-10}$ esu; most of the space within the atom is empty. Atomic nuclei are immutable through ordinary chemical changes; when one molecule of hydrogen (H_2) combines with one molecule of chlorine (Cl_2) to give two molecules of hydrogen chloride (HCl), the four nuclei (two hydrogen nuclei, or protons, of charge $+1e$ and two chlorine nuclei of charge $+17e$) are unchanged. It is redistribution of electrons between atoms which constitutes chemical combination. This is what valences of atoms control, and this is what a theory of valence must explain.

Atomic structure. Understanding of molecule formation requires an understanding of the electronic structure of atoms. According to N. Bohr, electrons in an atom move in orbits much like the orbits of planets about a sun, held to the nucleus by electrical attractions for it, prevented from falling into it by centrifugal forces. A special quantum effect is operative at the atomic level, however, which possesses no analogy in the motions of planets; not all orbits are possible for an electron, but only those for which the angular momentum of the electron as it moves about the nucleus is an integer multiple of $h/2\pi$, where $h = 6.63 \times 10^{-34}$ J is Planck's constant, and for which the energy is similarly quantized. Furthermore, not more than two electrons can move in one orbit at once. *See* ATOMIC STRUCTURE AND SPECTRA; ELECTRON CONFIGURATION.

When the consequences of these ideas are worked out, there actually emerges the periodic classification of the elements. To cover just part of the periodic table, occupation of orbits by electrons in the lighter atoms are shown in the **table**, where the symbol 2*p* stands for three distinct orbits of the same energy and shape but differently oriented in space. The lowest energy orbit is 1*s*, forming the K shell. Next in energy are 2*s* and 2*p*, making up the L shell. The 3*s* state is still higher, in the M shell. The chemically inert gases helium (He) and neon (Ne) are characterized by closed shells of 2, and $2 + 8 = 10$ electrons, respectively. The next inert gas is argon (Ar), with a closed shell of $2 + 8 + 8 = 18$ electrons, followed by krypton (Kr) with $2 + 8 + 18 + 8 = 36$ electrons, and the others. *See* INERT GASES; PERIODIC TABLE.

Electron configuration of some atoms

			Orbit					
			K shell	L shell		M shell		
Atom		*Z*	1*s*	2*s*	2*p*	3*s*	3*p*	3*d*
Hydrogen	(H)	1	1	0	0	0	0	0
Helium	(He)	2	2	0	0	0	0	0
Lithium	(Li)	3	2	1	0	0	0	0
Beryllium	(Be)	4	2	2	0	0	0	0
Boron	(B)	5	2	2	1	0	0	0
Carbon	(C)	6	2	2	2	0	0	0
Nitrogen	(N)	7	2	2	3	0	0	0
Oxygen	(O)	8	2	2	4	0	0	0
Fluorine	(F)	9	2	2	5	0	0	0
Neon	(Ne)	10	2	2	6	0	0	0
Sodium	(Na)	11	2	2	6	1	0	0
Magnesium	(Mg)	12	2	2	6	2	0	0
Aluminum	(Al)	13	2	2	6	2	1	0

Rule of eight. Many of the simple facts of valence follow from the postulate that atoms combine in such a way as to seek closed-shell or inert-gas structures (rule of eight) by the transfer of electrons between them or the sharing of a pair of electrons between them. Many molecular structures may be obtained by inspection by using these rules. The electrons in the K shell are not involved in the bonding for atoms after He, nor are the electrons in the K and L shells for atoms following Ne.

Hydrogen has a valence of 1, because one more electron will give a hydrogen atom an inert-gas structure. Carbon can form four bonds, because four more electrons give it the neon electronic structure.

Bond types. The bond between two atoms is covalent if one electron in the bonding electron pair comes from each atom, as in H:H or the CH bonds in CH_4. It is coordinate covalent if both electrons come from one atom, as the boron-nitrogen bond in the compound $F_3B{\rightarrow}NH_3$. If there is complete transfer of electrons from one atom to another, the bond is electrovalent or ionic, as in sodium fluoride (NaF). Bonds intermediate in type are possible; the bond in hydrogen fluoride (HF) is between covalent and ionic. An ionic bond X^+Y^- will be more stable the less the ionization potential of X and the greater the affinity of Y for electrons, that is, when X is a metallic element from the lower left corner of the periodic table and Y is a nonmetallic element from the upper right corner. Bond type can be inferred from both chemical and physical evidence. *See* CHEMICAL BONDING; ELECTRONEGATIVITY.

Bonds involving one or three electrons are known, but they are rare; H_2^+ and HeH are examples. Multiple bonds between atoms are common and important; examples are the carbon-carbon bond and the carbon-oxygen bonds in ethylene and carbon dioxide. For discussion of a bond of special importance in biology *see* HYDROGEN BOND.

Valence electrons are the electrons of an atom that can participate in chemical bonding, for example, for H and He the 1*s* electrons, for Li through Ne the 2*s* and 2*p* electrons, and for Na the 3*s* electrons.

Oxidation-reduction. As generally used, the word valence is ambiguous. Before a value can be assigned to the valence of an atom in a molecule, the electronic structure of the molecule must be known exactly, and this structure must be describable simply in terms of simple bonds. In practice, neither of these conditions is ever precisely fulfilled. A term not so ambiguous is oxidation number or valence number. Oxidation numbers are useful for the balancing of oxidation-reduction equations, but they are not related simply to ordinary valences. Thus the valence of carbon in CH_4, $CHCl_3$, and CCl_4 is 4; oxidation

numbers of carbon in these three substances are −4, +2, and +4. *See* OXIDATION-REDUCTION.

Quantum theory of valence. The above simple theory of valence is inadequate in at least three ways. First, it fails to account for many experimental facts, such as why the six C—C bonds in the molecule benzene, C_6H_6, are physically and chemically equivalent, what the electronic structures of the boron hydrides are, why the H—H bond is much stronger than the C—C bond, why CO_2 is a linear molecule but H_2O nonlinear, and what principles govern the rates of chemical combination. Second, the explanations that are offered are not physically satisfying. The stability conferred upon a molecule by the sharing of a pair of electrons by two atoms is established, but what is the real origin of this stability? And third, the theory is not detailed or quantitative enough to allow correlation and prediction of the many different properties of molecules. Dozens of properties of molecules can be measured, many to a high degree of accuracy. The ultimate theory should account for all of these quantitatively. *See* MOLECULAR STRUCTURE AND SPECTRA.

The quantum theory of valence does not possess these faults. It is based on the precise laws of physics for the atomic domain that were formulated in the 1920s by E. Schrödinger and others, the discipline known as quantum mechanics. The primitive quantum ideas of M. Planck and N. Bohr require modification to take care of the experimental fact that electrons and other particles at times act like waves. Like waves, they interfere when they are on top of one another in a manner that can be precisely calculated. According to nineteenth-century physics, an electron moving about a proton would collapse onto it. In the Bohr theory this collapse is prevented by a special quantum hypothesis; in the new mechanics it is prevented by elementary energy considerations. It would be favored by the attractive potential energy of the particle pair, but it turns out to be catastrophic for their kinetic energy. Instead of collapse a compromise is reached; the electron, or wave, is smudged out over a region surrounding the nucleus which defines the atomic size. *See* QUANTUM MECHANICS; UNCERTAINTY PRINCIPLE.

The pattern of the periodic table comes out as before. The orbits of Bohr are replaced by other entities, orbitals, which represent not the paths of the electrons but the amplitudes of the electron waves at different points in space. Furthermore, electrons are treated as if they were spinning, but only in two possible ways. The rule that generates the periodic table then is that in an atom no two electrons can occupy the same atomic orbital with the same spin. *See* EXCLUSION PRINCIPLE.

In a chemical bond, again there is interplay of kinetic and potential energies. An electron pair will tend to be shared by two atoms instead of being located on one of them if that situation is energetically favorable. The region between nuclei is more favorable for the potential energy of electron-nuclear attraction than other regions the same distance from just one nucleus. Moving in this region is not as favorable for the kinetic energy as moving on individual atoms, but the potential energy predominates when a bond is formed. The normal covalent bond may be described as two electrons occupying one molecular orbital, rather than two distinct atomic orbitals, with opposite spins because the exclusion principle is still operative. *See* MOLECULAR ORBITAL THEORY.

When a detailed examination is made of these effects with the modified theory, the stabilities of actual molecules and other of their properties are quantitatively accounted for. In particular, if two atoms approach which have low-energy atomic orbitals which overlap each other in space, and if two electrons are available, the conditions are favorable for forming, with evolution of heat, a chemical bond. It follows that the valence of an atom is given by the number of unpaired electrons it possesses, an old basic rule of valence.

The greater the overlap between two atomic orbitals, the stronger the bond that can be formed with them (criterion of maximum overlapping). This condition may be regarded as determining the shapes of molecules. Two or more orbitals of comparable energy, as 2*s* and 2*p* orbitals, can be combined (hybridized) to give orbitals concentrated along certain directions in space, and these are the orbitals that participate in directed bond formation. In the carbon atom, for instance, the four electrons in the 2*s* and 2*p* subshells are potential valence electrons. The two 2*s* electrons are paired, however, so that to make four bonds possible one of these must be promoted to a vacant 2*p* orbital. Four bonds than are possible, in various directions. Four equivalent bonds can be formed, tetrahedrally directed, as in CH_4. Three bonds in a plane and one other less strong one can be formed, as in $H_2C{=}CH_2$. In this manner L. Pauling and others have accounted for a multitude of phenomena in stereochemistry. *See* STEREOCHEMISTRY.

The peculiar bonding in benzene and other aromatic molecules has been explained, together with its consequences for chemical reactivity. The principles governing reaction rates have been formulated and applied. *See* BENZENE; DELOCALIZATION; RESONANCE (MOLECULAR STRUCTURE).

Research in valence theory has led to general and complete understanding of most aspects of molecular electronic structure, and has contributed toward acceptance of the language of modern physics as a proper language for chemistry. Considerable research in this field continues, however, because substances with new types of bonds are being synthesized constantly, and new physical methods for studying molecules are constantly revealing more intimate details of molecular structure which demand explanation. *See* CHELATION; CHEMICAL DYNAMICS; CONJUGATION AND HYPERCONJUGATION; COORDINATION CHEMISTRY; LIGAND FIELD THEORY; ORGANIC CHEMISTRY; ORGANOMETALLIC COMPOUND; STRUCTURAL CHEMISTRY. Robert G. Parr

Bibliography. H. Gray, *Electrons and Chemical Bonding*, 1964; D. J. Klein and N. Trinajstic (eds.), *Valence Bond Theory and Chemical Structure*, 1990; R. McWeeney (ed.), *Coulson's Valence*, 3d ed.,

1980; M. F. O'Dwyer, J. E. Kent, and R. D. Brown, *Valency*, 2d ed., 1989; L. Pauling, *The Nature of the Chemical Bond and the Structure of Molecules and Crystals*, 3d ed., 1960; L. Pauling and P. Pauling, *Chemistry*, 1995; A. N. Stranges, *Electrons and Valence: Development of the Theory 1900–1925*, 1982.

Valence band

The highest electronic energy band in a semiconductor or insulator which can be filled with electrons. The electrons in the valence band correspond to the valence electrons of the constituent atoms. In a semiconductor or insulator, at sufficiently low temperatures, the valence band is completely filled, and the conduction band is empty of electrons. Some of the high energy levels in the valence band may become vacant as a result of thermal excitation of electrons to higher energy bands or as a result of the presence of impurities. When some electrons are missing, the remaining ones may be redistributed among the energy levels within the valence band under an applied electric field, giving rise to an electric current. The net effect of the valence band is then equivalent to that of a few particles which are equal in number and similar in motion to the missing electrons but each of which carries a positive electronic charge. These "particles" are referred to as holes. *See* BAND THEORY OF SOLIDS; CONDUCTION BAND; ELECTRIC INSULATOR; HOLE STATES IN SOLIDS; SEMICONDUCTOR.

H. Y. Fan

Bibliography. C. Kittel, *Introduction to Solid State Physics*, 7th ed., 1996.

Value engineering

A thinking system (also called value management or value analysis) used to develop decision criteria when it is important to secure as much as possible of what is wanted from each unit of the resource used. The resource may be money, time, material, labor, space, energy, and so on. The system is unique in that it effectively uses both knowledge and creativity, and provides step-by-step techniques for maximizing the benefits from both. It promotes development of alternatives suitable for the future as well as the present. This is accomplished by identifying and studying each function that is wanted by the customer or user, then applying knowledge and creativity to achieve the desired function. Resources are converted into costs to achieve direct, meaningful comparisons. By using the methods of value engineering, 15 to 40% reduction in the required resources often results.

Application. Value engineering has applications in five broad areas: in design, purchase, and manufacture of products; in administrative groups, private or public, where the task is to achieve accomplishment through people; in all areas of social service work, such as hospitals, insurance services, or colleges; in architectural design and construction; and in development as well as research.

Approach–function study. The specific functions which the customer or user wants will be of two types: use function, which does something for the user; and esthetic function, which is selected because it pleases the user. Value engineering uses a planned approach for intensive and effective utilization of every applicable technique. It requires the development of sufficient skill in the application of enough techniques to bring into clear view a liberal number of value alternatives. Each function is systematically understood, identified, clarified, and named.

Value. Value is defined as the proper function for the least resources. Best value, then, is the attainment of the full function desired for the lowest cost achievable. The value of a function becomes a vital measure, being the lowest cost of securing the function. After arriving at this measure, high effort is made to achieve a function near its value.

The system is used to improve value in either or both of two situations: (1) The product or service as used or as planned may provide 100% of the functions the user wants, but lower costs may be needed. The system then holds those functions but achieves them at lower cost. (2) The product or service may have deficiencies, that is, it does not perform the desired functions or lacks quality, and so also lacks good value. The system aims at correcting those deficiencies, providing the functions wanted, while at same time holding the use of resources (costs) at a minimum.

Evaluation of function by comparison. Once the functions have been identified, clarified, understood, and specified, the following question must be answered: What is the lowest cost which, under the present conditions, would provide the described function? The answer is developed by comparisons to the past. These values are established by other valid comparisons, such as: How might an important portion of the function be accomplished, and what would that cost? How would that function be accomplished in a different industry or a different country, under very different conditions, and what would it cost? These values have to be compared with larger, smaller, and similar items or services and their costs. If there is no comparison, there can be no evaluation.

Often this task of creatively evaluating the function of itself brings a good answer to the problem. For example, the Navy was building 1000 landing crafts. One function was to "contain 200 gallons of gasoline." A noncombat life of 8 years was desired. The best quotation for each landing craft was $520. The function was evaluated by comparisons. How else could 200 gal be contained and what would each cost? Four 50-gal drums would have cost $25. A standard 250-gal oil tank, often used with oil burners in homes, cost $30. The $30 figure is selected as a base. Additional costs were considered for some

connections, piping, and perhaps coatings that would be needed, so the value of the function was selected to be $50. As a result of the evaluation the Navy elected to use four drums, separated into two groups of two, at a unit cost of $80. Thus, the cost for the job was $80,000 instead of $520,000.

Minimizing normal human negatives. The system, with its intense emphasis on functions, deep searches for knowledge, and constant effective creativity, built into step-by-step techniques, offsets many human traits which retard or prevent beneficial change. Some of the human negatives which act as retardants are the following: (1) Thoughts tend to follow habit patterns. (2) A decision to change, if proved wrong, may bring embarrassment. (3) Making a change may bring personal risk to the decision maker. (4) Decisions based upon sound general criteria often do not fit the specific instance. (5) The required good decision may be contrary to what is "normally" done. (6) Subjective coloring of attitudes of important people in the area makes good decisions difficult without sound objective data. (7) Decisions vital to profitability are often made by people not accountable for profits. (8) Obscure cause-and-effect relationships in some matters that create costs allow decisions to be made that will injure value. (9) While feelings are strictly personal, they may influence and often control decisions affecting value. (10) Most environments are hazardous to anything new, or to a change. Sound, objective alternatives developed by the value engineering problem solving system do much to overcome these human retardants.

Job plan. All minds must be "tuned" to work on exactly the same problem at exactly the same time. A job plan is thus developed in order to present the problems that will be faced and to establish the functions that should be accomplished. The job plan requires the following five steps: (1) gathering extensive information about the problem area; (2) analyzing the information for meaning and sense of direction; (3) doing the essential creative work; (4) judging the results; (5) creating a development planning program. The first four steps all require a different mental activity, and each has to be thoroughly completed before proceeding onto the next step.

Information step. In this step, all the facts and pertinent information (such as costs, quantities, and specifications) have to be obtained because only through complete understanding of the situation can valuable assessments be made. Assumptions have to be sorted out and reviewed to determine if they can be replaced or supported by facts. Long-standing assumptions have to be especially checked for validity.

Analysis step. This step involves the development of "function" thinking. From the information collected in the first step, functions are developed to answer such questions as: What are the problems involved, and which must be solved first? Are the solutions to these problems reasonable? What goals should be aimed for, and what steps should be taken in order to achieve them? Is any more information needed? Have any assumptions been overlooked, and are all assumptions already noted still valid? Have the best approaches to these problems been developed and, if so, what savings or benefits will result? Should better solutions be sought? What problems, if presented with better solutions, would produce even more beneficial results for the project?

Creativity step. After acquiring the relevant information and reaching an understanding of the problems through analysis, one must apply creative thinking, involving free use of the imagination. Two aspects of the mind are important here: the diverse knowledge bits that it has acquired over the years, and the ability to join these bits in various combinations constituting temporary "mental pictures" that may be unique. Thus a number of fresh alternative solutions to consider emerge in solving the overall problem or individual problems.

Judgment step. This step involves creatively studying the ideas presented in the previous step. No ideas should be thrown out. Rather, they should be developed and improved. Ideas involving monetary value should especially be studied closely and objectively in order to seek out their limitations and to try to lessen, overcome, or eliminate any negative aspects. It may be necessary to send certain ideas back to step one and to run them through the entire process again. Some ideas may have so many advantages that the lessening of their drawbacks may become the principal point of concern. And those ideas which appear to be completely thought through and seem capable of providing the greatest yield are sent on to step five.

Development planning step. In this step the best specialists and vendors are selected for consultation, and an investigation program is established to provide the most recent information on and the most current capabilities of any of the approaches that show potential.

VECP (value engineering change proposals). When this thinking system is used on government or military supply or construction situations, the same or better quality is usually secured for millions of dollars less in cost. To make it profitable for the manufacturer or contractor to hire value engineers and make this contribution to the government, a percentage of the saving, often 40 to 50%, for a time is paid on VECPs. Hundred of millions of tax dollars are being saved each year because of VECPs. In addition, resources are being conserved, and millions of dollars are being saved by owners, because of the work of the architectural firms and construction companies that have trained and qualified value engineers.

Training and qualification. Skilled consultants and some universities teach the techniques of value engineering. A professional society, the Society of American Value Engineers, with chapters in many cities, sets standards, gives examinations, and awards the citation of "Certified Value Specialist" (CVS) to qualified people.

World status. Extensive application of value engineering is growing in the United States, Japan, Germany, Sweden, France, Canada, and England. Important application is growing in Norway, Italy, Spain, Korea, Taiwan, South Africa, India, and other places. Professional societies exist in the United States, Japan, Scandinavia, France, and South Africa. *See* ACTIVITY-BASED COSTING; INDUSTRIAL ENGINEERING; METHODS ENGINEERING; OPERATIONS RESEARCH; OPTIMIZATION; PROCESS ENGINEERING; PRODUCTION ENGINEERING; PRODUCTION PLANNING.

Lawrence D. Miles

Bibliography. E. Castillo, *Extreme Value Theory in Engineering,* 1988; W. H. Copperman, *A Guide to the Contractual Aspects of Value Engineering,* 1986; R. M. Curtice, *Strategic Value Analysis: A Modern Approach to Systems and Data Planning,* 1987; M. C. Macedo, Jr., P. V. Dobrow, and J. J. O'Rourke, *Value Management for Construction*, 1978; A. Mudge, *Value Engineering: A Systematic Approach,* 1989; J. J. O'Brien, *Value Design and Construction*, 1976; D. E. Parker, *Management Application of Value Engineering: For Business and Government*, 1994; *Value Engineering and Management Digest*, monthly.

Valve

A flow-control device. This article deals with valves for fluids, liquids, and gases. Valves are used to regulate the flow of fluids in piping systems and machinery. In machinery the flow phenomenon is frequently of a pulsating or intermittent character and the valve, with its associated gear, contributes a timing feature. For electrical valves *see* ELECTRON TUBE

Pipe valves. The valves commonly used in piping systems are gate valves (**Fig. 1**), usually operated closed or wide open and seldom used for throttling; globe valves (**Fig. 2**), frequently fitted with a renewable disk and adaptable to throttling operations; check valves (**Fig. 3**), for automatically limiting flow in a piping system to a single direction; and plug cocks (**Fig. 4**), for operation in the open or closed position by turning the plug through 90° and with a shearing action to clear foreign matter from the seat.

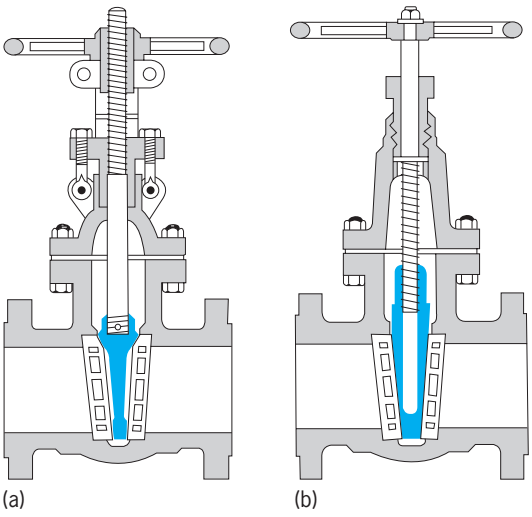


Fig. 1. Gate valves with disk gates shown in color. (*a*) Rising threaded stem shows when valve is open. (*b*) Nonrising stem valve requires less overhead.

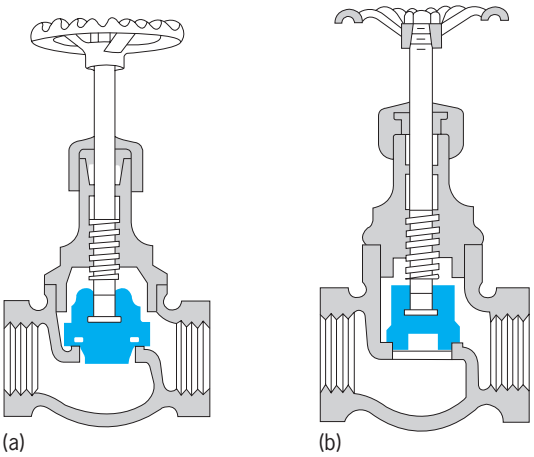


Fig. 2. Globe valves with (*a*) gasket in disk and (*b*) ground metal-faced disk.

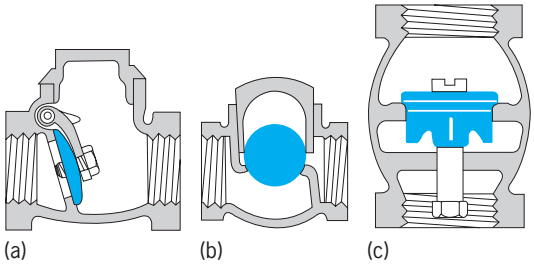


Fig. 3. Various types of straightway check valves. (*a*) Swing. (*b*) Ball. (*c*) Vertical.

Valves may have various structural features such as outside stem and yoke; packless construction; angle, as opposed to straightway flow; power instead of manual operation; and combined nonreturn and stop-valve arrangements. Valves are made in a wide assortment of materials, and a wide variety of trim, with brass or bronze for general service; cast iron for low steam pressures and temperatures (less than 250 lb/in.^{2} or 1.7 megapascals) and for hydraulic pressures below 800 lb/in.^{2} (5.5 MPa); steel and alloy steels for the highest operating pressures and temperatures (such as 5000 lb/in.^{2} or 34 MPa, 1200°F or 650°C steam); and selected metals for chemical and process applications. Most valves are manufactured and available as hardware and comply with the requirements of the ASTM, ANSI, and ASME as to material and dimensional standards. They are variously offered as flanged, screwed, welded, sweated, or compression-fitted for connection to pipe, machinery, and fittings.

Safety and relief valves are automatic protective devices for the relief of excess pressure. They are usually rigorously specified under the legal

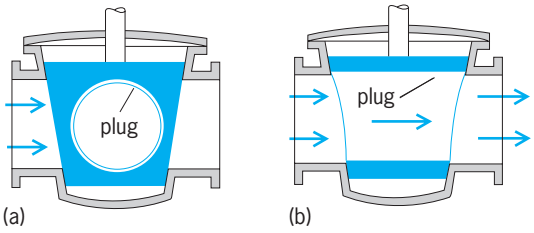


Fig. 4. Plug valve in (*a*) closed and (*b*) open positions.

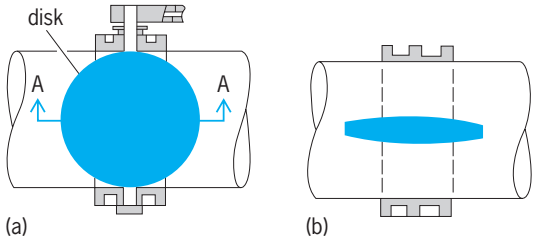


Fig. 5. Butterfly-type valve for penstock is typically 25 ft (7.6 m) in diameter. (*a*) Elevation. (*b*) Section A-A.

regulations of public authorities and insurance underwriters. They must open automatically when the pressure exceeds a predetermined value; they must allow the pressure to drop a predetermined amount before closing to avoid chattering, instability, and damage to the valve and the valve seat; they must have adjustment features for both the relieving and blowdown pressures; and they must be tamperproof after setting by responsible licensed operators. *See* SAFETY VALVE.

Hydraulic turbine valves. For hydraulic turbines and hydroelectric systems, valves and gates control water flow for (1) regulation of power output at sustained efficiency and with minimum wastage of water, and (2) safety under the inertial flow conditions of large masses of water. Valve sizes are usually large [for example, 6 ft (1.8 m) in diameter] so that power operation is necessary. Carefully streamlined construction to minimize fluid dynamic losses must be accompanied by ample provision to withstand shock and damaging effects of hydraulic inertia. Gate, butterfly (**Fig. 5**), telescoping, and needle constructions (**Fig. 6**) are variously employed. Wicket or cylinder gates regulate the flow of water to a reaction turbine at the speed ring while a governor-operated needle valve regulates flow to a Pelton impulse unit. *See* HYDRAULIC TURBINE; WATER HAMMER.

Steam-engine valves. To control the kinematics of the cycle, steam-engine valves range from simple D-slide and piston valves to multiported types. Slide valves control admission and release of steam to and from a double-acting cylinder by a single moving valve mechanism giving the necessary lap, lead, and angle of advance to accomplish the predetermined values of cutoff and compression. Multiported valves such as plug, Corliss, or poppet valves provide four valves for a double-acting cylinder. Each valve serves a single purpose of admission or exhaust for the head or crank end. The uniflow construction uses a poppet valve for admission with a row of exhaust ports alternately covered and uncovered by the engine piston. Plug and slide valves are limited to low pressures and temperatures (200 lb/in.^{2} or 1.4 MPa and 100°F or 55°C of superheat); poppet valves will operate under the maximum pressures and temperatures of steam-engine practice without warping. Many types of reversing gear have been perfected which use the same slide valve or piston valve for both forward and backward rotation of an engine, as in railroad and marine service. *See* STEAM ENGINE.

Internal combustion engine valves. Poppet valves are used almost exclusively in internal combustion reciprocating engines because of the demands for tightness with high operating pressures and temperatures. The valves (**Fig. 7**) are generally 2 in. (5 cm) in diameter or smaller on high-speed automotive-type engines, are cam-operated and spring-loaded, with their lift a small fraction of an inch, and with a gas velocity of 200–300 ft/s (60–90 m/s). They are cooled by transfer of heat to the engine jacket mostly through the valve stem. In heavy-duty units, the stem may have a partial mercury or sodium filling which aids, by inertia, in conducting heat to the

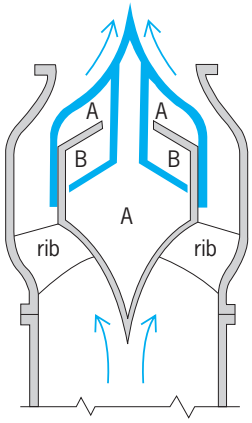


Fig. 6. Large needle valve is actuated hydraulically by pressure in chambers A to close and by pressure in annular chambers B to open.

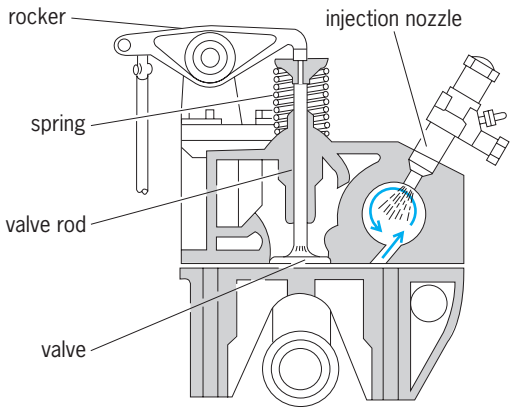


Fig. 7. Poppet valve for internal combustion engine. (*After E. A. Avallone and T. Baumeister III, eds., Marks' Standard Handbook for Mechanical Engineers, 8th ed., McGraw-Hill, 1978*)

valve stem and guides. Exhaust valves are subject to the effects of extreme temperature and must accordingly be most carefully designed and constructed of alloy materials. Two-cycle engines utilize ports, alternately covered and uncovered by the main piston, for inlet or exhaust. *See* CAM MECHANISM; INTERNAL COMBUSTION ENGINE; VALVE TRAIN.

Compressor valves. In compressors, valves are usually automatic, operating by pressure difference (below 5 lb/in.^{2} or 34 kilopascals) on the two sides of a movable, spring-loaded member and without any mechanical linkage to the moving parts of the compressor mechanism (**Fig. 8**). The practical objectives are timing, tightness, and life. Pressure drop should be a minimum. Moving parts must be light, but spring loadings should be heavy enough to avoid chatter and to give certainty to the valve action. Airspeeds are in the order of 5000 ft/min (1500 m/min) through the valves. Valves for inlet and exhaust service are frequently made interchangeable in the interest of reduced investment. On high-pressure (above 1000 lb/in.^{2} or 7 MPa) service, automatic poppet valves are usually substituted for the more common plate valves to give greater strength and tightness under loading. *See* COMPRESSOR.

Pump valves. Like those for compressors, pump valves are usually of the automatic type operating by pressure difference. The service conditions, however, are very dissimilar because of the noncompressibility of liquids; the presence of entrained solids like grit, fibers, and sludge; the corrosive potential of chemicals like acids and alkalies; the high viscosity of many liquids; the vapor pressure at the pump-operating temperature; and inertia effects accompanying discontinuity of the liquid column. Fluid speeds are low (200–300 ft/min or 60–90 m/min with cold water and 100 ft/min or 30 m/min with viscous liquids). A multiplicity of small valves (about 4 in. or 10 cm in diameter) arranged in docks with a positive suction is a common construction. Various constructions are used (**Fig. 9**). Theodore Baumeister

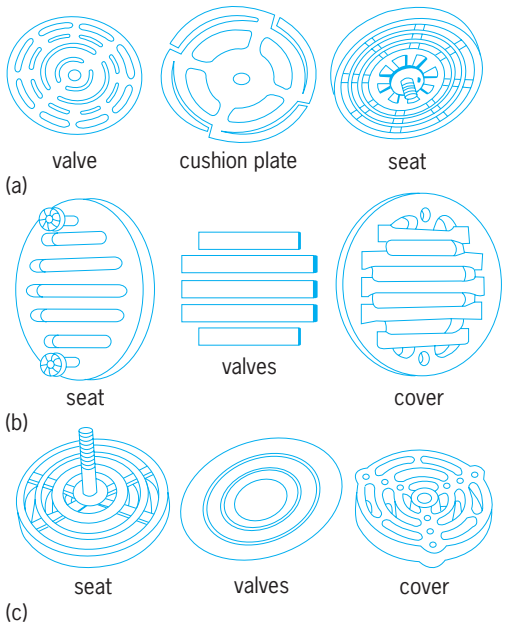


Fig. 8. Air and gas compressor valves. (*a*) Ingersoll-Rand ring plate. (*b*) Worthington feather valve. (*c*) Chicago pneumatic simplate valve. (*After E. A. Avallone and T. Baumeister III, eds., Marks' Standard Handbook for Mechanical Engineers, 8th ed., McGraw-Hill, 1978*)

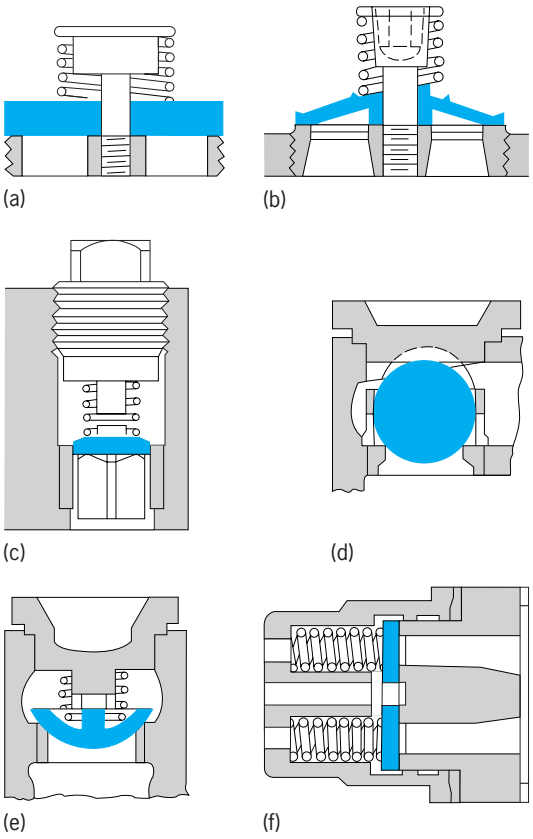


Fig. 9. Pump valves. (*a*, *b*) Disk valves are for moderate pressure. (*c*) Conical-faced wing valves are for high pressure (above 1000 lb/in.^{2} or 7 MPa). (*d*, *e*) Ball and rounded valves are for viscous liquids. (*f*) Double-ported valves are for high-speed reciprocating pump.

Bibliography. E. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; W. Ulanski, *Valve and Actuator Technology*, 1991.

Valve train

The valves and valve-operating mechanism by which an internal combustion engine takes air or fuel-air mixture into the cylinders and discharges combustion products to the exhaust. *See* VALVE.

Mechanically, an internal combustion engine is a reciprocating pump, able to draw in a certain amount of air per minute. Since the fuel takes up little space but needs air with which to combine, the power output of an engine is limited by its air-pumping capacity. The flow through the engine should be restricted as little as possible. This is the first requirement for valves. The second is that the valves close off the cylinder firmly during the compression and power strokes. *See* INTERNAL COMBUSTION ENGINE.

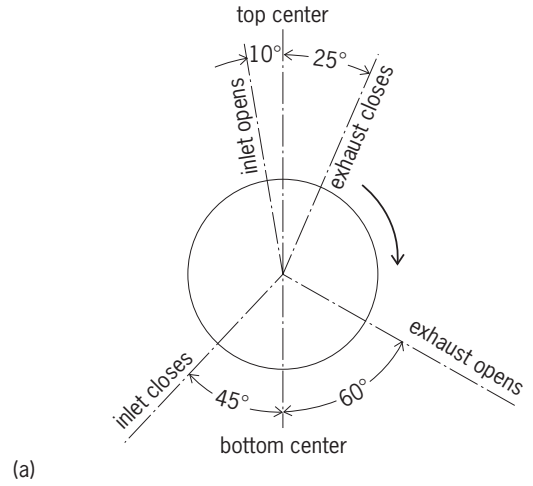


Fig. 1. Valve for a four-cycle engine. (*a*) Valve-timing diagram. (*b*) Valve train with the camshaft in the cylinder block and the valves in the cylinder head. (*Pontiac Division, General Motors Corp.*)

Valve action. In most four-stroke engines the valves are the inward-opening poppet type, with the valve head ground to fit a conical seat in the cylinder block or cylinder head (**Fig. 1**). The valve head is held concentric with its seat by a cylindrical stem running in a valve guide. The valve is held closed by a compressed helical spring. The valve is opened wide by lifting it from its seat a distance equal to approximately 25% of the valve diameter. At this lift the cylindrical opening between valve and seat will be about as large as the cross section of the flow passage to the valve seat.

The valve is streamlined and as large as possible to give maximum flow, yet of low inertia so that it follows the prescribed motion at high engine speed. Valves are usually made of a heat- and scale-resistant alloy that will keep its strength and shape at high temperature. The exhaust valve is sometimes made hollow and partially filled with metallic sodium. In operation the sodium melts and shakes back and forth, transferring heat from the valve head to the stem and valve guide. The valve-seat area is hard-faced, with inserts of a tough heat-resistant alloy used in aluminum castings and high-performance engines.

Cam action. Engine valves are usually opened by cams that rotate as part of a camshaft, which may be located in the cylinder block (Fig. 1) or cylinder head (**Fig. 2**). Riding on each cam is a cam follower or valve lifter, which may have a flat or slightly convex surface or a roller, in contact with the cam. The valve is opened by force applied to the end of the valve stem. A valve rotator may be used to rotate the valve slightly as it opens.

In many small four-stroke cycle engines, which have the camshaft and valves in the cylinder block, the force is applied through a valve lifter to the valve stem. In automotive and diesel engines with the camshaft in the cylinder block and valves in the cylinder head, the force may be applied through a mechanical linkage actuated by the cam follower (Fig. 1). The linkage consists of cam follower, pushrod, and rocker arm. The pushrod is a light rod or tube with ball ends, which carries the motion of the cam follower to the rocker arm. The rocker arm is a lever pivoted near its center so that, as the pushrod raises one end, the other end depresses the valve stem, opening the valve. *See* CAM MECHANISM.

In engines with the camshaft and valves in the cylinder head, the cam may operate the valve directly through a cup-type cam follower, or bucket tappet (Fig. 2). Other overhead-camshaft engines use a finger-type cam follower or rocker arm between the cam and the valve stem.

To ensure tight closing of the valve even after the valve stem lengthens from thermal expansion, the

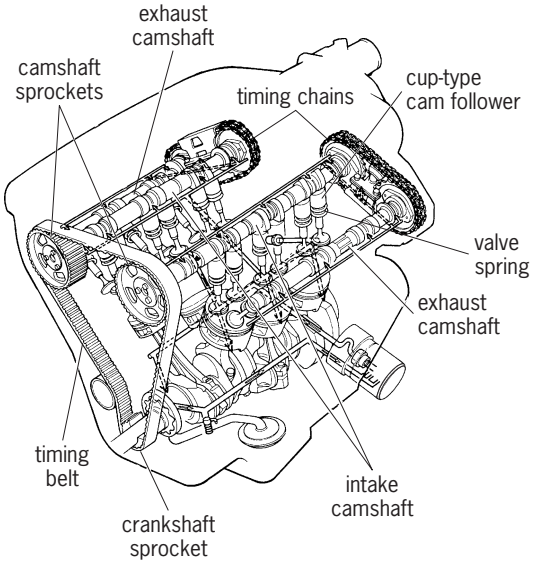


Fig. 2. Valve train of a V-6 four-cycle engine with the valves and two camshafts located in each cylinder head. The cams act directly on cup-type cam followers to open the valves. (*Ford Motor Co.*)

valve train is adjusted to provide some clearance when the follower is on the low part of the cam. The cam shape includes a ramp that reduces shock by starting the lift at about 2 ft (0.6 m) per second even though the clearance varies from time to time. To open the valve quickly, an acceleration of about 400 times gravity (for automotive-size engines) is used. Excessive acceleration deflects the valve linkage, giving false motion to the valve, causing it to close at high velocity before the closing ramp is reached. The cam surface between the opening and closing acceleration sections includes the point of maximum lift and maximum deceleration. If this part of the cam is sinusoidal and the valve spring is properly designed, the spring forces remain proportional to the forces required by the cam for the deceleration of the valve. The highest operating speed for a given lift can be obtained without the follower leaving the cam. The time that the valve is held at maximum lift is known as the dwell. *See* HYDRAULIC VALVE LIFTER.

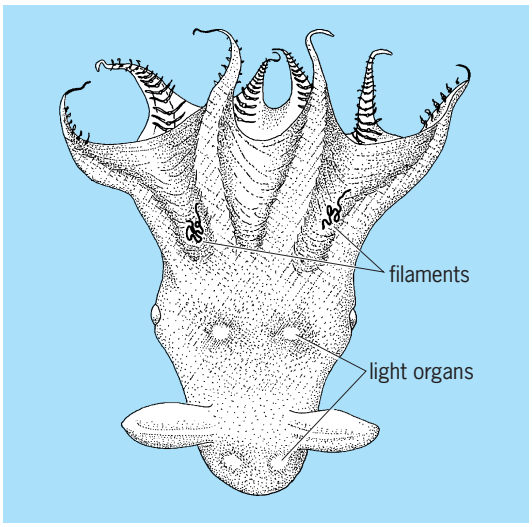
Timing. At high piston speeds the inertia of the charge in the inlet passages causes the gas to continue to flow into the cylinder long after the piston has started up on the compression stroke. To trap the maximum amount of fresh charge in the cylinder, the closure of the inlet valve may be delayed up to about 60° after bottom center. Late inlet closing increases engine torque at high speeds but reduces torque at low speeds. To reduce the work required at high speeds to expel the products of combustion during the exhaust stroke, the exhaust valve must open considerably before bottom center. The time during which both valves are open is known as valve overlap.

Donald L. Anglin

Bibliography. *Bosch Automotive Handbook*, 2d ed., 1987; W. H. Crouse and D. L. Anglin, *Automotive Engines*, 8th ed., 1994; Society of Automotive Engineers, *SAE Handbook,* 4 vols., annually.

Vampyromorpha

An order of coleoid cephalopods that contains only one species, *Vampyroteuthis infernalis* (see **illustration**). It is characterized by a flat, broad, leaflike, chitinous internal shell (gladius); eight arms around the mouth connected by a deep web; no tentacles, but two small sensory filaments that retract into pockets between the bases of the first and second arms; fingerlike cirri and a single row of suckers with chitinous rings along the arms; one pair of paddle-shaped fins on the body in adults (two pairs at the juvenile stage); very dark maroon, mostly black pigmentation; and photophores (light-producing organs) on the body, head, and arms. Vampyromorphs superficially resemble cirrate octopods, but they bear features that show distinct teuthoid (squid) relationships (for example, similar gladius, photophores, and radula), and in many characteristics are intermediate between teuthoids and octopods. Fossil records indicate the group is a descendant of Late Triassic or Jurassic forms. The sensory filaments between the first and second pair of arms are unique among all cephalopods and are thought to be tactile organs. *See* TEUTHOIDEA.



Vampyroteuthis infernalis **showing sensory filaments and photophores (light organs).**

The sexes are separate, and the male has no hectocotylus (modified arm for transferring sperm). Sperm are stored by females in seminal receptacles anterior to the eyes. The eggs are relatively large (0.12–0.16 in. or 3–4 mm in diameter), and they are laid individually in the water. The young usually occur deeper than 2900 ft (900 m), then gradually move shallower as they grow. Adults reach a total length of about 16 in. (40 cm). *See* CEPHALOPODA; COLEOIDEA; OCTOPODA.

Vampyroteuthis infernalis is a gelatinous bathypelagic (very deep, midwater living) species that inhabits worldwide tropical, subtropical, and temperate waters. It occurs mostly at depths of 1600–3800 ft (500–1200 m), occasionally to 4800 ft (150 m).

Clyde F. E. Roper

Bibliography. P. R. Boyle, *Cephalopod Life Cycles*, vol. 1, 1983, vol. 2, 1987; K. N. Nesis, *Cephalopods of the World*, transl. from Russian, 1987; G. Pickford, *Vampyroteuthis infernalis* Chun, an archaic dibranchiate cephalopod, pt. 1, Dana Rep. 29, pp. 1–45, 1946, pt. 2, Dana Rep. 32, pp. 1–132, 1949; C. F. E. Roper and R. E. Young, Vertical distribution of pelagic cephalopods, *Smithson. Contrib. Zool.*, no. 209, pp. 1–51, 1975; K. M. Wilbur and M. R. Clarke (eds.), *The Mollusca*, vol. 12: *Paleontology and Neontology of Cephalopods*, 1988; R. E. Young, Homology of retractile filaments of vampire squid, *Science*, 156(3782):1633–1634, 1967.

Vanadium

A chemical element, V, with atomic number 23. Natural deposits contain two isotopes, ^{50}V (0.24%), which is weakly radioactive, and ^{51}V (99.76%). Commercially important as an oxidation catalyst, vanadium also is used in the production of alloy steel and ceramics and as a colorizing agent. Studies have demonstrated the biological occurrence of vanadium,

1																	18
1 H	2											13	14	15	16	17	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg	3	4	5	6	7	8	9	10	11	12	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

especially in marine species; in mammals, vanadium has a pronounced effect on heart muscle contraction and renal function. *See* PERIODIC TABLE; TRANSITION ELEMENTS.

Very pure vanadium is difficult to prepare because the metal is highly reactive at temperatures above the melting point of its oxide (663°C or 1225°F) from which it is produced. Vanadium is a bright white metal that is soft and ductile. It has a melting point of 1890°C (3434°F), a boiling point of 3380°C (6116°F), and a density of 6.11 g/cm^3 (3.53 oz/in.^{3}) at 18.7°C (65.7°F). The thermal and electrical conductivity of vanadium is superior to that of titanium.

At room temperature, the metal is resistant to corrosion by oxygen, salt water, alkalies, and nonoxidizing acids, the exception being hydrogen fluoride (HF). Vanadium cannot withstand the oxidizing conditions presented by nitric acid or aqua regia. At elevated temperatures it will combine with most nonmetals to form oxides, nitrides, carbides, arsenides, and other such compounds.

The physical properties of vanadium are very sensitive to interstitial impurities. The strength varies from 30,000 lb/in.^{2} (200 megapascals) in the purest form to 80,000 lb/in.^{2} (550 MPa) in the commercial grade. The melting point is markedly altered by small impurities; vanadium containing 10% carbon has a melting point of 2700°C (4892°F).

Vanadium has a low fission neutron cross section. This property combined with the metal's excellent retention of strength at elevated temperatures has made its use in atomic energy applications attractive.

Carbon and alloy steels consume more than half the vanadium produced in the United States. Many plate, structural, bar, and pipe steels contain vanadium to enhance strength and toughness. The basis for the unique properties of these carbon alloy steels is the formation of vanadium carbide. These carbides are extremely hard and wear-resistant; they do not coalesce readily, but maintain a state of fine dispersion. Many large steel forgings contain vanadium in the range 0.05–0.15%; here vanadium acts as a grain refiner, and also improves the mechanical properties of the forgings. Tool steels are another large class of vanadium-containing steels; vanadium ensures the retention of hardness and cutting ability at the elevated temperatures generated by the rapid cutting of metals.

The production of ferrovanadium, an iron alloy, is very important since the primary commercial use of vanadium is in steel. Ferrovanadium is produced by aluminum or silicon reduction of V_2O_5 in the presence of iron in an electric arc furnace. The commonly practiced aluminum reduction is exothermic, so that little additional heat from the arc is required. Silicon processing requires a two-stage reduction to achieve efficient operation. *See* FERROALLOY; STEEL MANUFACTURE.

Vanadium compounds, especially V_2O_5 and NH_4VO_3, are excellent oxidation catalysts in the chemical industry. Processes that employ such catalysts include the manufacture of polyamides, such as nylon; sulfuric acid production by the contact process; phthalic and maleic anhydride syntheses; and various oxidations of organic compounds such as the conversion of anthracene to anthraquinone, ethanol to acetaldehyde, and sugar to oxalic acid. Vanadium pentoxide is used as a mordant in dyeing and printing fabrics and in producing aniline black for the dye industry. Vanadium compounds are used in the ceramics industry for glazes and enamels. A wide range of colors can be obtained with combinations of vanadium oxide, zirconia, silica, lead, tin, zinc, cadmium, and selenium. *See* CERAMICS; DYEING; MORDANT.

Vanadium has long been recognized as an essential element in biological systems; however, the role of the metal often is obscure. Tunicates accumulate vanadium to levels 1 million times greater than the surrounding seawater. This vanadium was once thought to act as an oxygen carrier but now is believed to be an oxidation catalyst that repairs damage to the polymeric, protective tunic of these animals. The first vanadium-dependent enzyme, vanadium bromoperoxidase, was isolated from brown, red, and green marine algae (for example, *Ascophyllum nodosum*); this enzyme catalyzes the bromination of a variety of organic molecules by using hydrogen peroxide and bromide. This activity may be the source of many important brominated compounds that potentially may be used as antifungal and antineoplastic agents. *See* ENZYME.

A variety of physiological effects in mammalian systems have been reported, the most significant being in cardiovascular and renal function. Vanadate causes constriction of veins in the kidney and can alter the retention and excretion of sodium and chloride ions. Cardiac effects of vanadium are species-specific, with observed increases (rabbit and rat) and decreases (guinea pig and cat) in heart muscle contractility. Because vanadate is a potent inhibitor of Na,K-ATPase in the laboratory, it has been suggested that this is the site of the metal's action. However, the physiology of vanadate is probably more complicated, since vanadium behaves as a hormone mimic by elevating intracellular calcium ion (Ca^{2+}) levels in a process that is poorly understood. *See* BIOINORGANIC CHEMISTRY.

Vincent L. Pecovaro

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; D. R. Lide, *CRC Handbook Chemistry and*

Physics, 85th ed., CRC Press, 2004; A. S. Tracey and D. C. Crans (eds.), *Vanadium Compounds: Chemistry, Biochemistry, and Therapeutic Applications*, 1998.

Van Allen radiation

The high-energy, charged particles that are trapped into orbits by the geomagnetic field, forming radiation belts that surround the Earth. The belts consist primarily of electrons and protons and extend from a few hundred kilometers above the Earth to a distance of about 8 R_e (R_e = radius of Earth = 6371 km = 3959 mi). James Van Allen and coworkers discovered them in 1958 using radiation detectors carried on satellites *Explorer 1* and *3*, and they are often referred to as the Van Allen belts.

Motion of charged particles. A charged particle under the influence of the geomagnetic field follows a trajectory that can be conveniently described as a superposition of three separate motions. The first motion, produced by the magnetic force acting at right angles to both the particle velocity and the magnetic field, is a rapid spiral about magnetic field lines. As the spiraling particle moves along the field line toward either the North Pole or South Pole, the increase in magnetic field strength causes the particle to be reflected so that it bounces between the Earth's two hemispheres. Superimposed on the spiral and bounce motions is a slow east-west drift; electrons drift eastward and protons or heavier ions drift westward. Thus, individual trapped particles move completely around the Earth in a complicated pattern, their motion being constrained to lie on magnetic shells. **Figure 1** illustrates the motion of a trapped proton during several north-south reflections. Subsequent bounces will carry the proton completely around the Earth. Trapped electrons behave in a similar manner except that the direction of drift is eastward. In the absence of perturbing forces such as additional electric and magnetic fields or collisions with other particles, this motion will continue indefinitely.

Because the Earth's magnetic field is not axially symmetric about the Earth, the radiation belts are also asymmetric. For example, the trapped particles dip closer to the Earth as they drift over the Atlantic Ocean. At altitudes above a few R_e, this effect becomes small. Near the outer boundaries of the belts, at 7–10 R_e, an additional warping of the geomagnetic field results from the pressure of the solar wind.

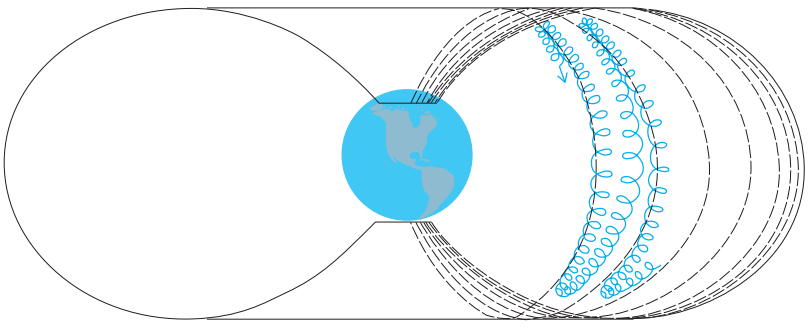


Fig. 1. Motion of a proton trapped in the geomagnetic field.

Trapped particle populations. The distribution of charged particles in the Earth's radiation belts is illustrated schematically in **Fig. 2**, where electron fluxes are depicted in the cross section on the left and proton fluxes are shown on the right. The darker shading indicates more intense fluxes of trapped particles. The spatial structure of trapped radiation shows two maxima: an inner radiation belt centered at about 1.5 R_e and an outer belt centered at 4–5 R_e. In the inner radiation belt, the most penetrating particles are protons with energies extending to several hundred megaelectronvolts (MeV). The intensity of the high-energy protons reaches 10^5 cm^{-2} s^{-1} for energies above 15 MeV. However, the flux of high-energy protons decreases rapidly with increasing distance from the Earth and becomes insignificant beyond 4 R_e. Low-energy protons with energies up to a few MeV occur throughout the stable trapping region which extends to approximately 8 R_e. Electrons are also found throughout the trapping region, with local maxima occurring in the inner and outer belts. The electron energies extend to several MeV, and in the outer radiation belt electrons are the most penetrating component.

Experiments have shown that helium nuclei (alpha particles) with energies of a few MeV are also trapped in the geomagnetic field, the maximum flux of a few hundred particles per square centimeter per second (energy >2 MeV) occurring at about 3 R_e. The flux of alpha particles is therefore only about 2×10^{-4} that of the protons. Large fluxes of trapped oxygen ions with energies up to about 20 kiloelectronvolts (keV) have also been discovered. These oxygen-ions, which are most pronounced after magnetic storms, are important in that they imply that some of the trapped ions come from the Earth's atmosphere.

The **table** gives sample values of the omnidirectional, integral fluxes of trapped electrons and protons as a function of altitude (in R_e) for positions in the equatorial plane. These values are averages, and time variations can be substantial, particularly for electrons in the outer belt. Also given in the table are values of the characteristic energy E_0, for which the flux decreases by a factor of 2.7. The table shows the systematic decrease in E_0 with increasing altitude for protons. A similar trend occurs for electrons except within the region between the inner and outer belts where the electron flux exhibits a minimum and the energy spectrum is richer in low-energy electrons.

Earth's ring current. The trapped particles drift slowly around the Earth. This motion, caused by inhomogeneities in the Earth's magnetic field, causes electrons to drift eastward and protons or heavier ions to drift westward. The resulting current, called the ring current, acts to decrease the strength of the Earth's (surface) northward magnetic field at low latitudes. During times of magnetic storms, it can store significant amounts of energy. To a first approximation, the ring current particles and radiation belt

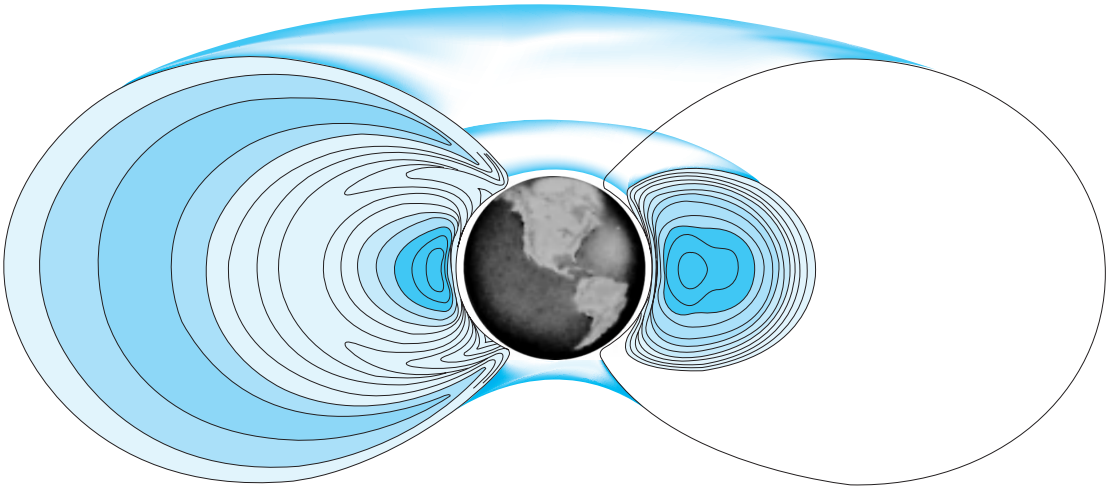


Fig. 2. Structure of Earth's radiation belts. Protons ($E > 15$ MeV) are shown on the right, and electrons ($E > 0.5$ MeV) are shown on the left. Darker shading indicates more intense fluxes of trapped particles.

particles are the same. There are, however, some important distinctions. Radiation belt particles typically refer to those particles that can penetrate into dense materials, such as instruments on board spacecraft or even astronauts. Electrons, for example, are important sources of penetrating radiation, yet contribute little to the ring current. Conversely, ring current particles carry the bulk of the total current density. *See* CURRENT DENSITY.

Time variations. The intensity, energy spectrum, and spatial distribution of particles within the radiation belts vary with time. The most dramatic variations are associated with magnetic storms, although smaller changes accompany most magnetic fluctuations. The changes are most pronounced for particles in the outer belt where the magnetic variations are the largest. During a magnetic storm, the electron flux in the outer belt may increase by an order of magnitude or more, although the response differs greatly among individual storms. Following a magnetic storm, the electron fluxes at lower altitudes usually increase as electrons, which were accelerated into the outer belt, diffuse closer to the Earth.

Figure 3 illustrates the time variations in the flux of electrons above 0.5 MeV on the Equator at 4 R_e. The upper part of the figure gives the K_p index (a measure of the magnetic activity). The higher the values of K_p, the larger are the irregular fluctuations in the geomagnetic field. It is apparent that the electron flux increases noticeably following magnetic fluctuations and decays slowly while the Earth's field is steady. The fluxes of low-energy protons in the outer belt also increase during magnetic storms, although the changes in proton fluxes are generally smaller than for electrons.

In the inner belt, the high-energy proton flux is relatively constant with time such that a very strong magnetic storm is required to alter the intensity appreciably. The most pronounced effect observed during magnetic storms is a reduction in the high-energy trapped proton flux, the reduction being greater with increasing distance from the Earth.

At the outer edges of the radiation belt, large time variations are common. A systematic diurnal variation exists, caused by the warping of the overall geomagnetic field by the pressure of the solar wind (a stream of ionized plasma emitted by the Sun). The resulting asymmetry in the field produces a diurnal variation in the trapped radiation observed at a given point above the Earth. Variations in solar-wind pressure alter the degree of warping of the magnetic field and produce variations in the flux of charged particles. *See* GEOMAGNETIC VARIATIONS; MAGNETOSPHERE; SOLAR WIND.

Omnidirectional fluxes and characteristic energies of trapped particles on the Equator

Altitude, R_e	Low-energy protons ($E > 0.4$ MeV)		High-energy protons ($E > 15$ MeV)		Electrons ($E > 0.5$ MeV)	
	Flux, $cm^{-2}\ s^{-1}$	E_0, MeV	Flux, $cm^{-2}\ s^{-1}$	E_0, MeV	Flux, $cm^{-2}\ s^{-1}$	E_0, MeV
1.25	4×10^4	2.2	1×10^4	20	4×10^7	0.9
1.5	1×10^6	2.2	1×10^5	13	6×10^7	0.4
2.0	1.6×10^7	2.1	7×10^4	8	3×10^6	0.13
2.5	1×10^8	0.6	8×10^3	8	2×10^5	0.33
3.0	2×10^8	0.4	2×10	7	6.7×10^4	0.42
3.5	1×10^8	0.34			3×10^5	0.5
4.0	3×10^7	0.3			1×10^6	0.4
5.0	4×10^6	0.13			2.5×10^6	0.35
6.0	9×10^5	0.11			9×10^5	0.25

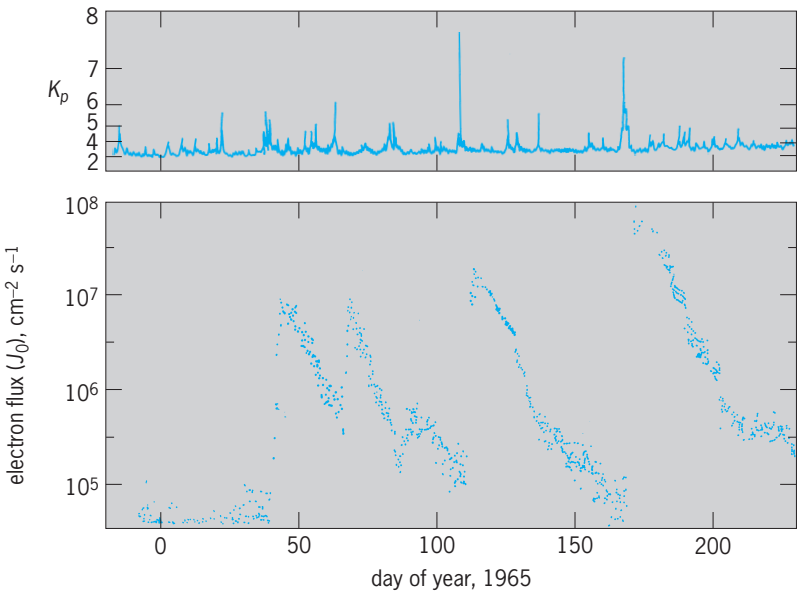


Fig. 3. Time variations in the magnetic activity index K_p and the associated changes in the flux of trapped electrons ($E > 0.5$ MeV) at an altitude of 4 R_e.

Artificial radiation belts. On at least nine occasions, high-energy electrons were injected into the geomagnetic field and trapped as a result of high-altitude nuclear detonations. The most famous of these events was the Starfish (hydrogen bomb) test on July 9, 1962, which produced an intense radiation belt lasting for many years. Following this nuclear explosion, several satellites ceased to transmit data because of loss of electrical power resulting from solar-cell damage. Shorter-lived belts were induced by three nuclear tests conducted by the Soviet Union in the fall of 1962. *See* NUCLEAR EXPLOSION.

Particle accelerators that are carried on rockets and satellites can also inject trapped electrons and ions. This type of experiment is designed to study the characteristics of the geomagnetic field and the plasma instabilities, which affect a beam of charged particles.

Origin and loss of trapped particles. It is believed that most of the very high energy protons (>50 MeV) in the inner belt result from the spontaneous decay of high-energy neutrons, produced by collisions of cosmic rays with atmospheric atoms. However, the vast majority of the radiation belt populations are ions and electrons that originate from either the atmosphere or the solar wind and are accelerated by processes only partly understood.

Theories of the sources of radiation-belt particles call on fluctuating electric and magnetic fields to accelerate the trapped particles. The geomagnetic field is continually agitated by variations in solar-wind pressure, and the changing magnetic field can accelerate charged particles by means of the induced electric fields. In addition to accelerating charged particles, the fluctuating fields also cause the particles to migrate across magnetic shells, making it possible for a particle that was initially injected at the outer boundary of the geomagnetic field to be transported to the inner belt. Detailed calculations of the acceleration and diffusion processes resulting from magnetic variations are not yet possible since there is insufficient knowledge about electric and magnetic fields in space. As knowledge improves, it will be possible to provide more exacting tests of the theories.

The physical nature of the processes by which trapped particles are ultimately lost from the belts is also poorly understood. The average lifetime for electrons varies from about a year in the center of the inner radiation belt to a few days in the outer belt. Within about 1000 km (600 mi) from the Earth, the particles are removed by collisions with neutral atoms and molecules of the Earth's atmosphere. The cumulative effect of these collisions is that the particles eventually lose so much energy that they become part of the ionosphere. At higher altitudes, the loss mechanism is probably the interaction of the trapped particles with electromagnetic waves. Waves of many frequencies and propagating modes exist throughout the magnetosphere, and their electric and magnetic fields can disrupt the trapping motion of the particles, eventually allowing them to reach the lower atmosphere. Theoretical estimates of the loss rates produced by these effects are in fair agreement with experimental values, but the severe approximations, which are required in the calculations, make the results inconclusive. Experiments have shown that electrons are also removed from the radiation belts by electromagnetic waves from high-power radio transmitters on the ground and by lightning discharges; however, the importance of these processes in controlling the radiation belts remains controversial.

Relation to other phenomena. Trapped radiation plays an important role in magnetic storms. One of the most important features of a magnetic storm is the main phase, during which the electric currents resulting from the motion of the trapped electrons and ions cause the horizontal component of the Earth's field measured at the Equator to decrease.

Electrons and protons, which are guided down magnetic field lines to the atmosphere at hight latitudes, also excite the polar aurora. It is believed that some of the same processes that accelerate auroral particles also accelerate radiation-belt particles. *See* AURORA.

The radiation belts are just one feature of the space plasma environment. For example, plasma from the Sun is continually impinging on the Earth's magnetic field, resulting in a "cavity" known as the magnetosphere. In addition, ionospheric electrons and ions form a plasma. Some of the electrons and ions making up this plasma can be accelerated and become trapped in the geomagnetic field, forming the radiation belts.

Particles are continually leaking out of the radiation belts into the atmosphere. This process produces ionization and contributes to the electron and ion population in the ionosphere. It is likely that some of the energy necessary to produce the airglow is derived from the charged particles striking

the atmosphere. The precipitating particles also initiate chemical reactions in the upper atmosphere that can, in turn, alter the concentrations of some of the minor species. For example, intense precipitation of electrons or protons can reduce the ozone content of the upper atmosphere, although it is not clear whether the effect is important on a global scale. *See* AIRGLOW; IONOSPHERE.

Radiation belts of other planets. Because the conditions that lead to the formation of Earth's radiation belts are so general, it is believed that any planet or moon that has a large enough magnetic field will also have radiation belts. Spacecraft missions to Mercury, Venus, and Mars have shown that the magnetic moments of these planets are too small to support long-term trapping of energetic particles. However, Jupiter, Saturn, Uranus, and Neptune have strong magnetic fields and very large, intense radiation belts analogous to those of the Earth. The belts for these planets have many characteristics that differ from those of the Earth because these planets are farther from the Sun, have high rotation rates, and have moons and rings immersed in their radiation belts. The rings of Saturn are efficient absorbers of energetic particles, so the radiation belts have gaps in the region of the rings. Similarly, the numerous moons of Saturn and Jupiter absorb particles, leading to minima in the radiation belt profiles at the orbital distances of the moons. On the other hand, some moons, especially Io at Jupiter, have atmospheres from which atoms escape, are ionized, and become part of the radiation belts. Hence, the radiation belts of some outer planets contain heavier ions such as oxygen, sulfur, and sodium provided by moons. Other sources of particles, such as protons from decaying neutrons, protons and helium ions from the solar wind, and ions from the planets' atmospheres, apparently exist in much the same fashion as with Earth. Comparison of the radiation belts at the different planets is giving valuable insight into the importance of the various plasma processes that form this trapped population of particles. *See* PLASMA (PHYSICS).

Space weather and radiation belts. The term "space weather" describes the conditions in space that affect Earth and its technological systems. It is a consequence of the behavior of the Sun, and the interaction of the solar wind with the Earth's magnetic field. Space weather can have many adverse effects, including damage to critical spacecraft instrumentation, degradation in the performance of navigation and communications systems, disruption of power grids, and exposure of astronauts and even aircraft passengers to hazardous radiation levels. In particular, the fluxes of electrons and protons trapped in the radiation belts can injure both personnel and equipment on board spacecraft if the vehicle is exposed to these energetic particles for a sufficiently long time. For example, on January 11, 1997, the communications satellite *Telstar 401*, which was orbiting 22,000 mi above Earth, was lost during a magnetic storm caused by a coronal mass ejection hitting the Earth's magnetosphere. Because of the spatial structure of the belts, the degree of damage will be strongly dependent on the position in space and hence on the orbit of the vehicle. In the inner belt region, high-energy protons can pass through several centimeters of aluminum structure and injure components in the interior of the spacecraft. In most other regions of the radiation belts, the trapped particles are less penetrating, and damage is confined to exposed equipment such as solar cells. Pete Riley; Martin Walt

Bibliography. G. Beynon et al. (eds.), *The Magnetosphere, the High-Latitude Ionosphere, and Their Interactions*, 1989; Y. Kamide and W. Baumjohann, *Magnetosphere-Ionosphere Coupling*, 1993; M. G. Kivelson and C. T. Russell (eds.), *Introduction to Space Physics*, 1995; G. K. Parks, *Physics of Space Plasmas*, 1991; J. G. Roederer, *Progress in Solar-Terrestrial Physics*, 1983; M. Schulz, *Earth's radiation belts, Rev. Geophys. Space Phys.*, 20:613–621, 1982; M. Schulz and L. J. Lanzerotti, *Particle Diffusion in the Radiation Belts*, 1974; J. A. Van Allen, *Origins of Magnetospheric Physics*, 1983.

Van der Waals equation

An equation of state of gases and liquids proposed by J. D. van der Waals in 1873 which takes into account the nonzero size of molecules and the attractive forces between them. He expressed the pressure p as a function of the absolute temperature T and the molar volume $V_m = V/n$, where n is the number of moles of gas molecules in a volume V:

$$p = \frac{RT}{V_m - b} - \frac{a}{V_m^2}$$

Here $R = 8.3145$ J K^{-1} mol^{-1} is the universal gas constant, and a and b are parameters that depend on the nature of the gas. Parameter a is a measure of the strength of the attractive forces between the molecules, and b is approximately equal to four times the volume of the molecules in one mole, if those molecules can be represented as elastic spheres. The equation has no rigorous theoretical basis for real molecular systems, but is important because it was the first to take reasonable account of molecular attractions and repulsions, and to emphasize the fact that the intermolecular forces acted in the same way in both gases and liquids. It is accurate enough to account for the fact that all gases have a critical temperature T_c above which they cannot be condensed to a liquid. The expression that follows from this equation is $T_c = 8a/27\,Rb$. *See* CRITICAL PHENOMENA; GAS; INTERMOLECULAR FORCES; LIQUID.

In a gas mixture, the parameters a and b are taken to be quadratic functions of the mole fractions of the components since they are supposed to arise from the interaction of the molecules in pairs. The resulting equation for a binary mixture accounts in a qualitative but surprisingly complete way for the many kinds of gas-gas, gas-liquid, and liquid-liquid phase

equilibria that have been observed in mixtures. *See* PHASE EQUILIBRIUM.

The equation is too simple to represent quantitatively the behavior of real gases, and so the parameters *a* and *b* cannot be determined uniquely; their values depend on the ranges of density and temperature used in their determination. For this reason, the equation now has little practical value, but it remains important for its historical interest and for the concepts that led to its derivation. *See* THERMODYNAMIC PRINCIPLES.

J. S. Rowlinson

Bibliography. A. Ya. Kipnis, B. E. Yavelov, and J. S. Rowlinson, *Van der Waals and Molecular Science*, Oxford University Press, 1996.

Vanilla

A choice flavoring obtained from a climbing orchid, *Vanilla fragrans*, a native of tropical American forests. The vanilla plant belongs to the orchid family and is indigenous to southeastern Mexico, where it was used by the Aztecs to flavor their cocoa. In 1510 *vagnuila* first appeared in Spain. Its fruits are pods called vanilla beans (see **illus.**). These are picked at the proper time before they have fully matured.

Vanillin (4-hydroxy-3-methoxybenzaldehyde) is the principal component of vanilla, although other components contribute to the distinctive flavor of the extract compared to synthetic vanilla. When they are harvested, the beans contain no free vanillin; it develops during the curing period from glucosides that break down during the fermentation and sweating of the beans. The sweating process consists of alternately drying the beans in sunlight and bunching them so that they heat and ferment. Sweating boxes are used in Mexico, whereas the shorter Madagascar method starts out by wilting green pods in hot water and uses blankets on which the beans can first be spread out and later rolled up for the enzymatic reactions and fermentation to take place. Further curing and dehydration occur in a warehouse. Periods of 4 weeks to 4 months may be required to develop the proper flavor and reduce the moisture content of the beans sufficiently to prevent molding. Beans can be artificially dried in ovens, but frequently an inferior-quality product results. *See* FERMENTATION; FOOD ENGINEERING.

Pods and flowers of *Vanilla fragrans*. (*USDA*)

After curing, the pods are sorted into grades based on quality. The best cured beans are 8–10 in. (20–25 cm) long, with drawn-out ends and curved bases. They are soapy or waxy to the touch, dark brown, and coated with fine crystals of vanillin, termed frost. Vanillin constitutes 1.2–3.5% of the bean, but other compounds contribute also to the aroma. In addition to the flavoring materials, vanilla beans contain fat, wax, sugar, gum, resin, and tannin. Vanilla is used in cookery, confectionery, and beverages. Vanilla extract, most used, is prepared by extracting the crushed beans with alcohol. A synthetic vanillin is made from eugenol occurring in clove oil, but the natural product is preferred. Several plants have been used as substitutes for true vanilla but these are of little value. *See* ORCHIDALES; SPICE AND FLAVORING.

Perry D. Strausbaugh; Earl L. Core

The principal types of commercially used vanilla beans are the Mexican, Bourbon [Bourbon comes mainly from Madagascar, but was named after the island of Bourbon (now Island of Réunion) in the Indian Ocean, where the French started the cultivation of vanilla], South American, Javan, and Tahitian.

Vanilla extract is prepared from vanilla beans with or without one or more of the following added: sugar, dextrose, glycerol. Vanilla extract contains the soluble matters from not less than 3.3 oz of vanilla beans in 1 qt (10 g/100 ml). To be legally called vanilla extract, 1 U.S. gal (3.785 liters) of vanilla extract must contain the soluble matter from not less than 13.35 oz (378.5 g) of vanilla beans. The finished flavoring should contain at least 35% alcohol by volume to keep the solubles in solution.

In the alcoholic extraction of the vanilla flavor, the color of the extract is influenced by the quality of the beans, the strength of the alcoholic menstruum, the duration of the extraction, and the presence of glycerin, which is added to retard evaporation and to retain the flavor of the extract. Best results are obtained with three consecutive extractions at room temperature, each requiring a minimum of 5 days. The first should have a maximum alcohol content of 65%; the second, 35%; the third, 15%. To improve aroma, extracts are aged, using stainless steel or glass containers.

A standard vanilla extract is equivalent in flavoring strength, though not in quality, to a 0.7% vanillin solution. The vanillin content of pure extracts range from 0.04 to 0.12 oz/qt (0.11 to 0.35 g/100 ml), with the average at about 0.06 oz/qt (0.19 g/100 ml). Ash content, soluble ash, lead number, total acidity, and acidity other than vanillin are among the conventional indices used to detect adulteration.

The approximate average maximum levels of usage (in ppm) for vanilla extract in foods are as follows:

Beverages	200	Baked goods	1900
Ice cream, ices	3000	Syrups	54
Candies	4000	Icings	4800
Toppings	2700		

Gideon E. Livingston; Myron Solberg

Bibliography. H. B. Heath and G. A. Reineccius, *Flavor Chemistry and Technology*, 1986; R. Teranshi, R. G. Buttery, and F. Shahadi (eds.), *Flavor Chemistry: Trends and Developments*, 1989.

Vapor condenser

A heat-transfer device that reduces a thermodynamic fluid from its vapor phase to its liquid phase. The vapor condenser extracts the latent heat of vaporization from the vapor, as a higher-temperature heat source, by absorption in a heat-receiving fluid of lower temperature. The vapor to be condensed may be wet, saturated, or superheated. The heat receiver is usually water but may be a fluid such as air, a process liquid, or a gas. When the condensing of vapor is primarily used to add heat to the heat-receiving fluid, the condensing device is called a heater and is not within the normal classification of a condenser. *See* HEAT TRANSFER.

Classification by use. Condensers may be divided into two major classes according to use: those used as part of a processing system, and those used for serving engines or turbines in a steam power plant cycle.

Process condensers. In a processing system condensers selectively recover liquid from a mixed vapor, recover pure liquid from the vapor of an impure liquid, recover noncondensables from a mixture of gas and vapor, or extract heat from heat-pump, refrigeration, and cryogenic cycles. The vapor condensed may be steam or the vapor from any nonaqueous liquid. Condensers used for any purpose except as vapor condensers in the power plant cycle, even though they provide the heat sink for a process cycle, are classified as process condensers.

One of the more important uses of process condensers is that of condensate recovery. Condensers are frequently used with fractionating and distillation columns in the production of hydrocarbon liquids and in processes for liquefying gases. Similarly they are used in the production of distilled water as distiller condensers in single- and multiple-effect evaporating systems. The multistage flash evaporator with integral vapor condensers has been established as one of the more economical means for the production of potable water from seawater. This apparatus contains a multiplicity of condensers in series, each condenser producing a successively lower pressure in which flashed vapor from seawater brine is produced, condensed, and collected. Both distillate and the flashing brine streams reduce in temperature as they flow through the apparatus (**Fig. 1**).

Power cycle condensers. Used as part of the power plant cycle, condensers serve as the heat sink in the cycle and reduce the back pressure on the turbine or engine so that a maximum of heat energy becomes available as useful work. Steam (water) is the most common substance used for the power generation cycle. Other thermodynamic fluids can be used but seldom are. Another important function of a condenser in the steam power plant cycle is the recovery of condensate as the major source of pure boiler feedwater for the boilers. *See* BOILER FEEDWATER; THERMODYNAMIC CYCLE.

Classification of operation. Condensers may be further classified according to mode of operation as surface condensers or as contact condensers.

Surface condensers. In surface condensers the condensing vapor and the cooling fluid remain separated from each other by a dividing wall or walls, which form the heat-transfer surface. This heat-transfer surface is most often in the form of tubes but may also be plates or partitions of various geometries. Condensate, cooling fluid, and noncondensable gases are usually removed separately, although in some designs the condensate and noncondensable gases are withdrawn as a mixture. *See* SURFACE CONDENSER; VACUUM PUMP .

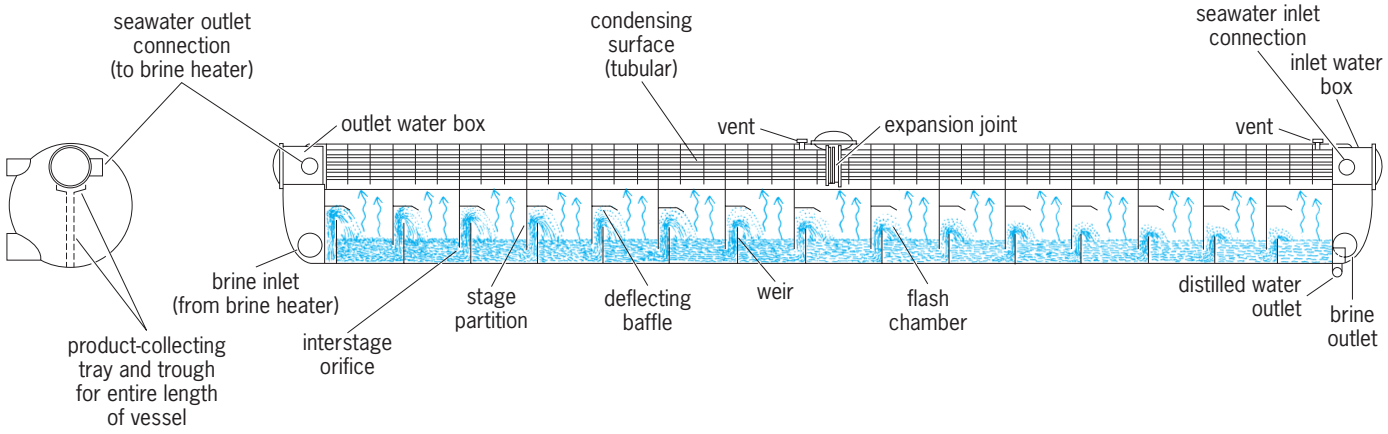


Fig. 1. Simplified views of multistage flash evaporator for production of potable water from seawater.

Contact condensers. In the contact condenser the vapor and cooling liquid come into direct contact with each other and are mixed in the condensing process. The condensed vapor and cooling liquid combine and are withdrawn from the apparatus together. The noncondensable gases are usually withdrawn separately, although in some contact condensers they are entrained in the mixed condensate and cooling liquid and removed through a common outlet. *See* CONTACT CONDENSER.

Removal of noncondensables. Condensers are required, almost without exception, to condense impure vapors, that is, vapors containing air or other noncondensable gases. Because most condensers operate at subatmospheric pressures, air leaking into the apparatus or system becomes a common cause for vapor contamination. Many process condensers are used to condense vapors whose noncondensable impurities are relatively independent of air that may leak into a vacuum system. Because the vapor supplied to the condenser is continuously reduced to liquid, the noncondensable gases in the gas-vapor mixture collect and concentrate. Their accumulation seriously affects heat transfer, and means must be provided to direct them to a suitable outlet. Most surface and contact condensers are arranged with a separate zone of heat-transfer surface within the condenser and located at the outlet end of the vapor flow path for efficient removal of the noncondensable gases through dehumidification. A vapor flow path of the displacement type, free from zones of stagnation and short circuiting, is essential for achieving maximum condensing heat-transfer rates (**Figs. 2** and **3**).

Separate external vapor condensers arranged in series with the vapor flow path of the main condenser are used when the ratio of noncondensable gases to condensing vapor is high, or when the vapor content of the noncondensable gases must be reduced to low values.

Removal of noncondensables from condensers operating at subatmospheric pressure requires vacuum pumps. The noncondensables are removed from condensers operating above atmospheric pressure by venting to atmosphere or to aftercondensers to reduce further the moisture content of the removed gases before discharging them to atmosphere or to a recovery process.

Heat-receiving fluid. Water is the most commonly used liquid for absorbing heat from condensing vapors. Liquid hydrocarbons and other chemical compounds in liquid state are used primarily as heat receivers for condensers, used as a part of a petroleum-refining or chemical-manufacturing process. Because of natural evaporation, surface waters are normally at a lower temperature than ambient air when air temperatures exceed 32°F (0°C), and thus provide the lowest-temperature cooling medium readily available. High specific heat, ease in pumping, and rapid heat transfer characterize water as an excellent medium for use as a heat receiver in a condensing system. The tendency of water to corrode metal surfaces is not usually a serious disadvantage. Thin-walled heat-transfer surfaces of the less costly corrosion-resistant alloys are generally satisfactory. Ferritic materials used for the water-containing parts of condensers can be made suitable for the corrosion environment by increasing thickness as a corrosion allowance or by the use of corrosion-resistant linings or by cathodic protection. *See* CORROSION.

Air is the gas most commonly used for absorbing heat from a condensing process. However, its low heat capacity, low density, and relatively low heat transfer preclude it from being an ideal fluid for this purpose. Its availability and the fact that air does not readily corrode or foul heat-transfer surfaces offer some advantages in its use. Shortages of water for cooling or for use as a heat sink and the danger of thermal pollution of surface water have markedly

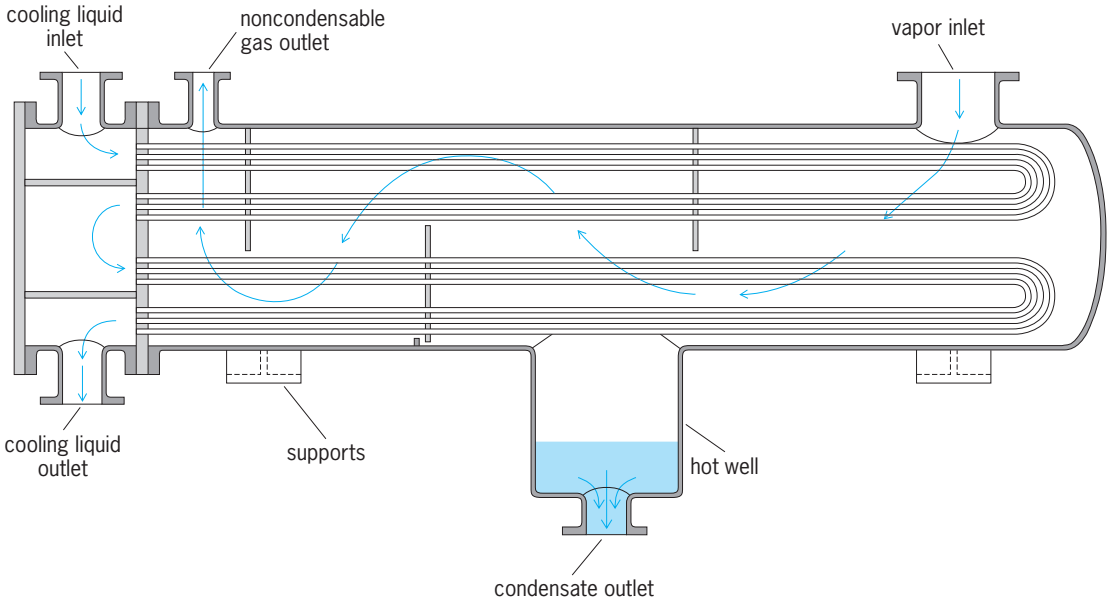


Fig. 2. Process condenser with one baffled-shell pass of condensing vapor and four tube passes of cooling liquid.

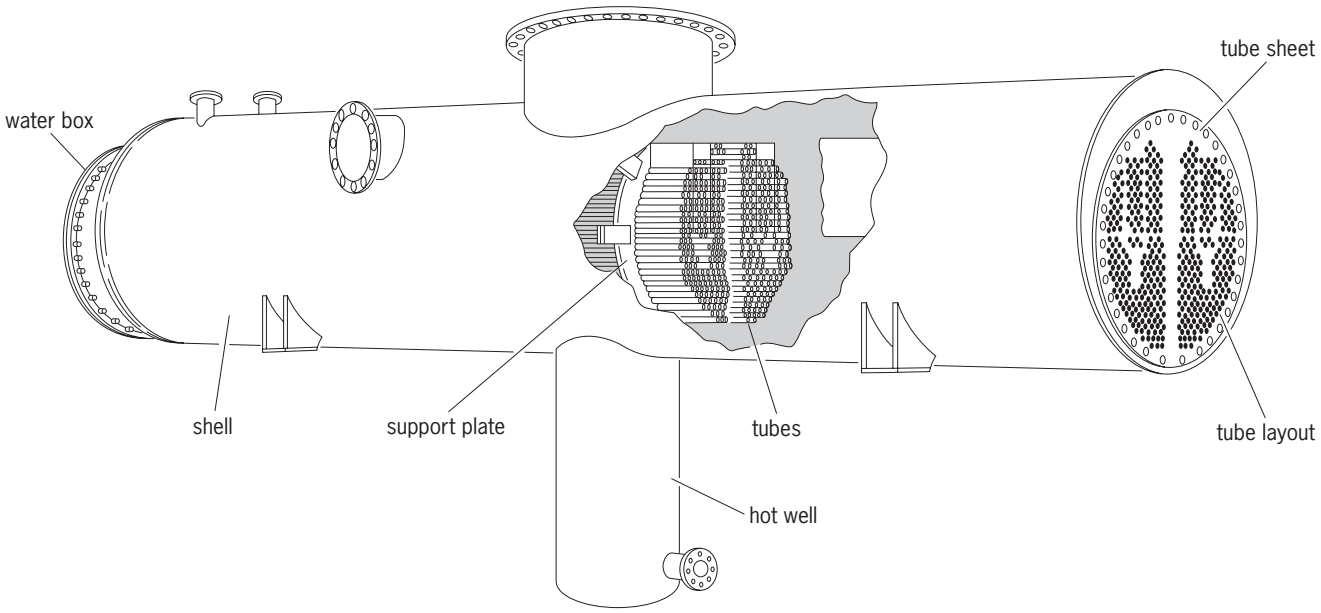


Fig. 3. Typical sections of small steam surface condenser.

increased the use of air. Application of extended-surface tubes to process condensers, to process coolers, and to steam condensers serving turbines is reasonably economical and practical where water supplies are critical.

The cooling of heat-transfer surfaces by evaporation involves the use of both water and air. The heat-transfer surfaces are continuously wetted by water. Air, blown over the wetted surfaces, absorbs the water vapor as it is released from the evaporating cooling water, thereby removing the heat released in condensers.

Condensate cooling. In those process systems where recovery of condensate is of primary importance, it is also desirable to cool the condensate below the saturation temperature of its condensing vapor. Some condensate cooling results as condensate falls from tube to tube in the process of being collected and accumulated. If additional cooling is needed, condensate cooling sections in the condenser can be provided by flooding a selected amount of surface within the condenser with condensate. Flooding is done either by installing baffles in the condensate drain zone to flood the tubes in these sections or by effecting a similar result with a loop in the condensate drain piping. Baffles are most effective in process condensers of the horizontal type, and loops are more adaptable to process condensers of the vertical type.

Condensate reheating. Modern steam surface condensers for turbines are provided with means for condensate reheating (**Fig. 4**). Condensate falling from tube to tube becomes cooled below the temperature corresponding to condenser operating pressure; unless provision is made for reheating this condensate before it is removed from the condenser, a measurable amount of heat energy is lost from the cycle, and cycle efficiency is reduced. In addition, subcooled condensate absorbs air from the condensing vapor and, if allowed to remain in the boiler feedwater, may cause serious corrosion of the feed system and boilers.

The condensate is reheated by a portion of the incoming steam. Baffles direct this steam to the condenser hot well in such a manner that its velocity

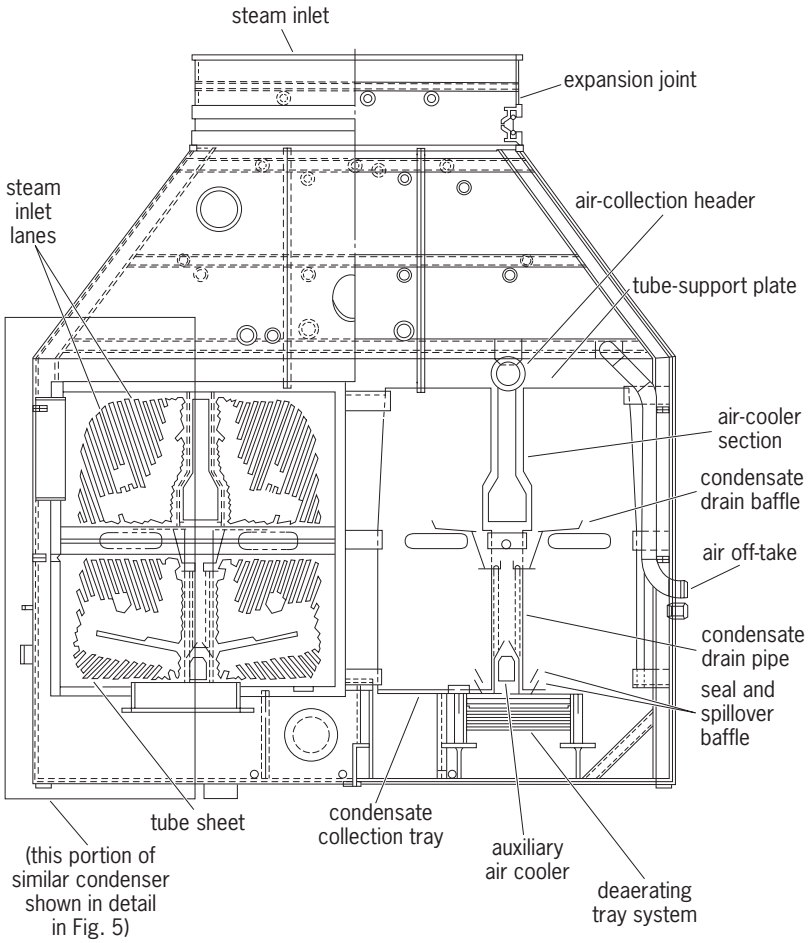


Fig. 4. Cutaway end view of steam surface condenser. (*Worthington Corp.*)

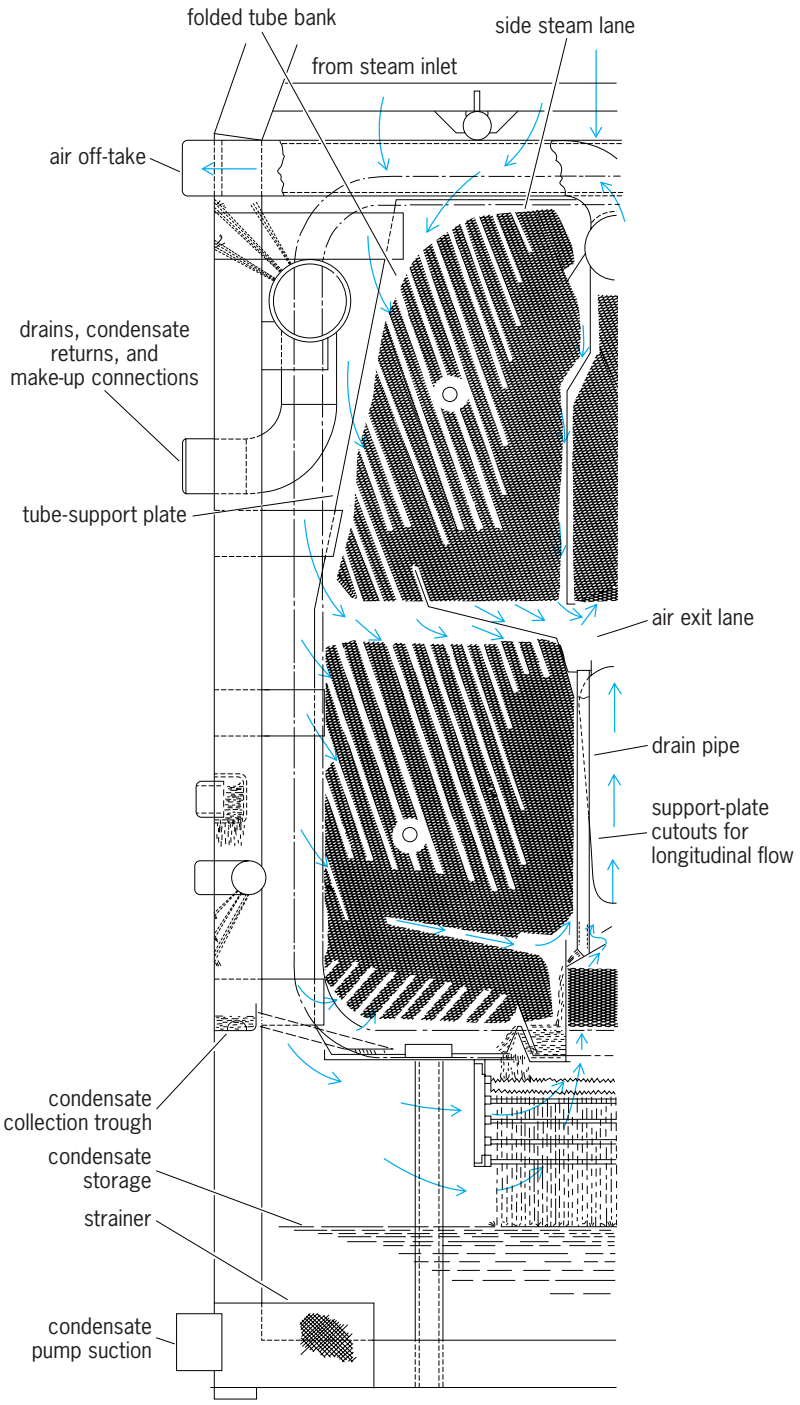


Fig. 5. Detail of a portion of a condenser that is similar to the one in Fig. 4. (*Worthington Corp.*)

energy is converted to pressure (**Fig. 5**). As a result, the local static pressure and corresponding temperature become equal to, or greater than, the static pressure and corresponding temperature at the condenser stream inlet. Condensate falling from the tube bundle to the hot well is reheated, the degree of reheat depending on the height of fall and the flow rate per unit area. If vertical height is limited, effective reheating and the associated deaeration can be achieved by redirecting the condensate falling from the tubes over collecting baffles and a series of trays or plates. This higher zone of pressure and temperature is vented along with the noncondensables removed from the condensate into the condenser tube bank and finally to the air cooler joining the mainstream of noncondensables being cooled and expelled. Under favorable conditions the condensate may reach temperatures as much as 5°F (3°C) above the temperature corresponding to static pressure at the condenser steam inlet, and the dissolved oxygen content of the deaerated condensate may be consistently less than 0.01 cm^3 per liter.

Condenser capacity. The quantity of cooling fluid required to condense a given quantity of vapor may be obtained by applying the general heat-balance equation. For a condensing vapor which may be superheated, saturated, or wet, with its condensate subcooled or reheated and with the noncondensable gases cooled, the relation is given by Eq. (1). Here

$$W_v(h_v - h_c) + W_n(h_{na} - h_{nb}) = W_f(h_{fb} - h_{fa}) \quad (1)$$

W_v = quantity of vapor; h_v = enthalpy of entering vapor; h_c = enthalpy of leaving condensate; W_n = quantity of noncondensable gases; h_{na} = enthalpy of entering noncondensables; h_{nb} = enthalpy of leaving noncondensables; W_f = quantity of cooling fluid; h_{fa} = enthalpy of entering cooling fluid; and h_{fb} = enthalpy of leaving cooling fluid. (The standard engineering practice is to express quantities in lb/h and enthalpies in Btu/lb.)

The temperature of the leaving cooling fluid is usually selected to be 5–10°F (3–6°C) less than the condensing temperature of the saturated vapor. Ordinarily, low temperature differences are associated with contact condensers; larger temperature differences are characteristic of economic design for surface condensers. Small temperature differences require enormous amounts of condensing surface and are seldom a practical design criterion.

For both contact and surface condensers the conditions for determining the hourly heat transferred can be expressed by Eq. (2) for undirectional steady-

$$Q = UA\ \Delta t_m \quad (2)$$

state heat flow. Here Q = hourly heat, Btu/h; U = overall heat-transfer coefficient, Btu/(h)(ft^2)(°F); A = area of heat-transfer surface, ft^2; and Δt_m = logarithmic mean-temperature difference, °F. Δt_m is based on the assumption of constant specific heat and substantially constant condensing temperature with the cooling-fluid temperature increasing from inlet to outlet in its flow path. U and A are generally applicable to surface condensers only.

For direct-contact condensers where surface is not readily determined, it is usual to consider heat transferred per unit volume of condensing space rather than per unit area. In this case Eq. (3) describes the

$$Q = KV\ \Delta t_m \quad (3)$$

heat flow, where K = volumetric heat-transfer coefficient, Btu/(h)(ft^3)(°F); V = volume of direct contact

condensing space, ft^3; and the other symbols are as previously defined.

The magnitude of K depends on the physical properties of the condensing vapor and heat-receiving fluid, the type of direct-contact surface (spray, tray, or packing), and the operating temperature. It is ordinarily determined experimentally for each type of contact condenser.

The magnitude of U, for use in Eq. (2), can be obtained from the summation of Eq. (4) for the individ-

$$U = \frac{1}{r_v + r_l + r_f + r_w} \tag{4}$$

ual resistances in the heat-transfer system, resistance being expressed in (h)(ft^2)(°F)/Btu. In Eq. (4), $1/r$ = conductance across any boundary, Btu/(h)(ft^2)(°F); r_v = condensing boundary resistance; r_l = cooling-fluid boundary resistance; r_f = fouling, dirt, or scale resistance; and r_w = separating wall, resistance.

Conductance of the condensing boundary $1/r_v$ can be determined analytically for pure vapors. For steam on the outside of horizontal tubes, $1/r_v$ may be 2000–2500 Btu/(h)(ft^2)(°F) [11,400–14,200 J/(s)(m^2)(°C)] for well-designed multitube condensers. Conductance of the cooling-fluid boundary $1/r_l$, for fluids flowing inside of tubes, may be determined from the relation of the Nusselt number to the product of the Reynolds and Prandtl numbers with the use of empirical constants such as those suggested by W. H. McAdams. Excellent agreement between computed values and test values results from the use of these equations. For water inside of tubes $1/r_l$ may be in the order of 1000–1400 Btu/(h)(ft^2)(°F) [5700–8000 J/(s)(m^2)(°C)] for the normal range of water velocities.

Values for conductance of the condensing boundary and the cooling-fluid boundary for some common thermodynamic fluids are listed in the **table** for cooling-fluid velocities in the range of 6–7 ft/s (1.8–2.1 m/s), condensing temperatures in the range of 100°F (38°C), and condensing rates of approximately 10 lb/(ft^2)(h) [1.4×10^{-2} kg/(m^2)(s)]. *See* CONDUCTION (HEAT); CONVECTION (HEAT).

Conductance for fouling or dirt films $1/r_f$ depends on the characteristics of the cooling fluid and the condensing vapor in relation to the accumulation (of dirt and corrosion products on the separating wall or tube surfaces) in service. In addition, the oxide film on new clear tubes or other types of metal heat-transfer surfaces contributes to overall fouling resistance r_f. For water-cooled shell and tube steam condensers a conductance of 2000 Btu/(h)(ft^2)(°F) [11,400 J/(s)(m^2)(°C)] is reasonable for copper-base alloy tubes after mechanical cleaning. Conductance of the separating wall $1/r_w$ may be computed from the thickness of the wall and the thermal conductivity of the material from which it is made. The overall heat-transfer coefficient varies with the physical properties, flow rates of the fluids, and geometry of the condenser and condensing surfaces. For steam condensers U may be in the order of 350–800 Btu/(h)(ft^2)(°F) [2000–4500 J/(s)(m^2)(°C)]; with few exceptions, heat transfer is lower with condensing vapors other than steam.

Conductance values for some thermodynamic fluids*

Fluid	Condensing vapor conductance	Cooling fluid conductance
Isopropyl alcohol	400	360
Benzene	600	520
Water	2500	1400
Ammonia	3200	2300

*Btu/(h)(ft^2)(°F) = 5.8 J/(s)(m^2)(°C).

Condenser components. The components of a representative contact condenser are shown in **Fig. 6**. The cooling-liquid distribution system shown consists of baffles and impingement surfaces to distribute the liquid uniformly and to allow it to cascade in counterflow relationship with the condensing vapor. In the illustrated condenser the cooling section for noncondensable gases coincides with the coolant distribution section where vapor is condensed and the noncondensable gases are dehumidified and concentrated. Other constructions for distributing the coolant include rings or slats made of metal, plastic, or ceramic. Most process condensers of the contact type are of counterflow design. In jet condensers noncondensables are entrained in the cooling liquid and thereby removed. The tail pipe provides the barometric leg and discharges into a sealing well, thus eliminating the need for an extraction or tail pump. Low-level contact condensers, those without barometric legs, usually require pumps to extract the cooling water.

Components of typical surface condensers are shown in Figs. 2, 3, 4, 5, and **Fig. 7**. The condensing surface consists of tubes of 0.5–1.5-in. (13–38-mm) outside diameter made of copper base alloys, less frequently of aluminum, nickel alloys, chromium steel, chromium-nickel steels, and titanium. The cooling surface for noncondensables is ordinarily the same material as the main condensing surface, although special corrosion-resistant alloys may be used when the noncondensables are especially corrosive. Usual practice is to condense the vapors on the outside of tubes (Figs. 2, 3, 4, and 5), but when the vapors are especially corrosive they may be condensed on the inside of tubes (Fig. 7). Tubes are usually plain, but for condensing vapors from low-conductance fluids, finned tubes may be used. For this purpose low fins, approximately equal in height to basic tube wall thickness, are used on the condensing side spaced so that the condensate formed does not bridge the fins and act like insulation. Gas- or air-cooled condensers use high fins on the cooling side with the extended surface generally at least 10 times the equivalent plain tube area. Bimetal tubes are most frequently used in process condensers when the corrosion environment is severe and different on one side of the tube than on the other.

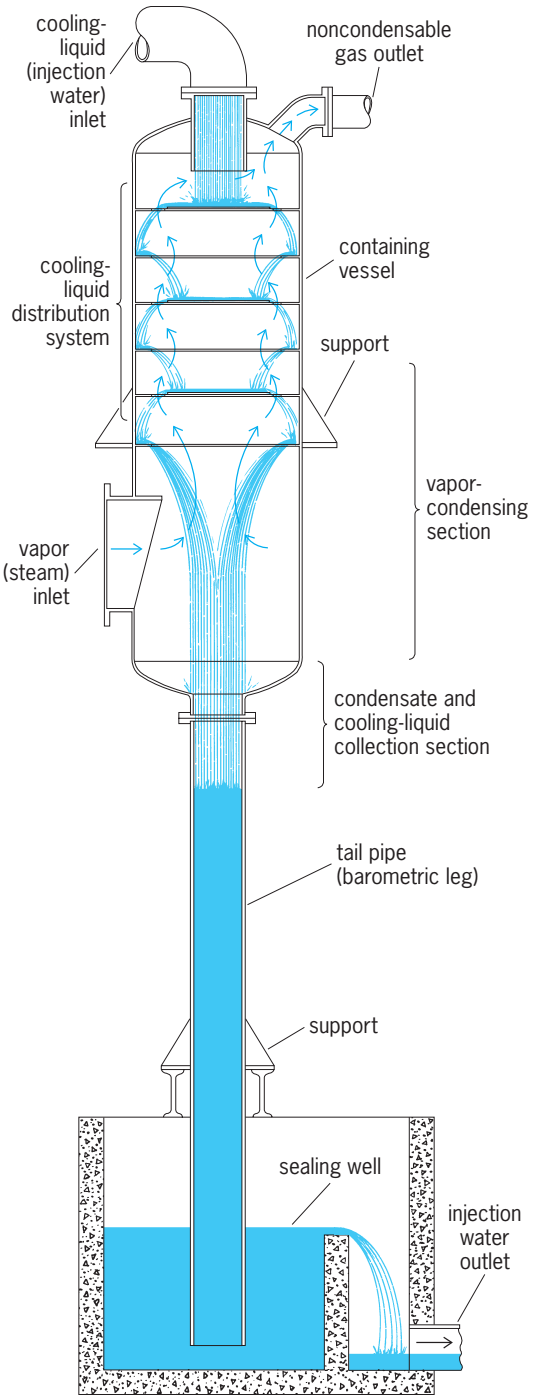


Fig. 6. Counterflow barometric-type contact condenser.

Tube vibration of magnitudes sufficient to cause tube failure is controlled by the spacing of tube support plates or baffles. Tube vibration seldom results from transmitted mechanical energy but rather from vapor velocity, causing tubes to deflect and vibrate at their natural frequency. With low-density vapors at high velocity, generally in the sonic range or greater, tube vibration is caused by the lift and drag effects of the vapor flowing around the tubes. Severe vibration is usually independent of the Kármán vortex street with low-density vapors, but may be associated with both the lift and drag effects of velocity and the Kármán vortex street at high-vapor densities even at relatively low velocities. *See* VORTEX.

The tube sheets (into which tubes are welded, expanded, or packed) are generally of Muntz metal, naval brass, copper-silicon alloy, or other corrosion-resistant alloys. Carbon-steel tube sheets may be used with carbon or chromium-nickel steel tubes. In nuclear power plants steam condenser tube sheets may be double, that is, with an air gap between to detect leakage and to prevent cooling water from leaking into the steam space. The inner tube sheets are usually made of carbon steel and the outer sheets are made of copper alloy.

The cooling fluid system for surface condensers, when liquid is used, usually consists of chambers attached to the tube sheets and arranged with inlet and outlet connections for the circulation of the cooling medium. When this fluid is liquid, the chambers are designed to distribute the liquid over the face of the tube sheets and into the tubes with little or no cavitation.

Condensers may be designed with one or more liquid passes, dependent on the thermal design conditions, the quantity of cooling liquid available or

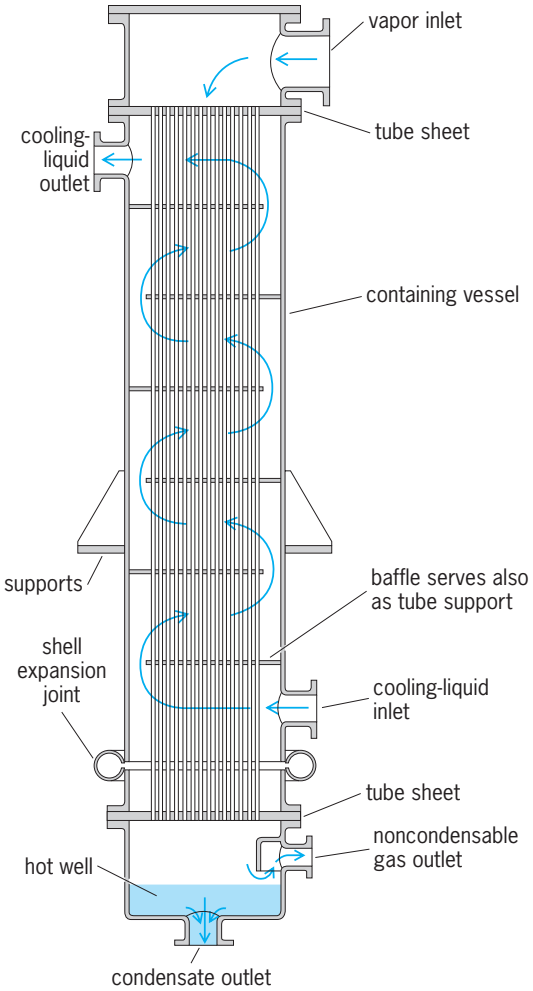


Fig. 7. Vertical-process surface condenser with condensing vapor inside the tubes and the cooling liquid outside the tubes, within the shell.

desired, and the space conditions for installation. Steam surface condensers in large steam power plants are of single-pass design except where the cooling-water supply is limited, in which case two-pass condensers are used. Large installations (those that require cooling towers, spray, or evaporation ponds as a means for controlling cooling-water temperature) ordinarily use two-pass condensers. Condensers with three or more passes seldom prove economical for serving engines or turbines. Process condensers are frequently designed with more than two cooling liquid passes and seldom with less.

Auxiliary equipment. Operation of condensers requires pumps (1) for injecting or circulating the cooling fluid, (2) for removing condensate or mixed condensate and injection water, and (3) for removing noncondensables.

Centrifugal pumps are usually used for injecting or circulating cooling water. Conventional volute pumps are used for circulating cooling water for small surface condensers and as injection water pumps for contact condensers. Mixed-flow volute pumps and axial-flow pumps of vertical design are best suited to large surface condensers.

Centrifugal pumps of the horizontal volute type, equipped with pressure-sealed stuffing boxes, are well suited for removing condensate. Vertical, multistage condensate pumps are more suited to large steam power plant installations, where they effect significant installation economies. Positive displacement pumps are used for withdrawing condensate from small condensers. Tail pumps of the conventional double-suction centrifugal type are used for pumping mixed condensate and injection water from low-level jet (contact) condensers. *See* PUMPING MACHINERY.

Pumps for removing noncondensable gases are classified as displacement and ejector types. The displacement machine is built either as a reciprocating vacuum pump, similar to a piston-type air compressor, or as a rotary machine, similar to a gear-pump or sliding-vane rotary compressor. The displacement-type vacuum pump is widely used on condensers of all sizes. It is not suitable for use at high vacuum with high noncondensable gas loads because it must be disproportionately large for such conditions. It is suitable for use with extremely large surface condensers serving turbines, because of the low noncondensable gas loads characteristic of these installations.

Steam jet ejectors are widely used as vacuum pumps. Having no moving parts, they require little or no operating attention and are simple to install. They have excellent capacity characteristics at high vacuum and are especially suited to applications where the noncondensable gas load is high. When used with steam condensers serving turbines, steam jet ejectors are equipped with surface inter- and aftercondensers. The exhaust steam is condensed and returned to the feed system, and the heat from the exhaust is also recovered in the feedwater. Thus they become highly efficient machines for removing noncondensable gases from condensers used in the generation of power. *See* CENTRIFUGAL PUMP; COOLING TOWER; DISPLACEMENT PUMP.

Joseph F. Sebald

Bibliography. A. J. Chapman, *Heat Transfer*, 4th ed., 1984; J. P. Holman, *Heat Transfer*, 9th ed., 2001; L. Iev and I. Iev, *Heat Transfer Flow Dynamics*, 1990; J. R. Simonson, *Engineering Heat Transfer*, 2d ed., 1989.

Vapor cycle

A thermodynamic cycle, operating as a heat engine or a heat pump, during which the working substance is in, or passes through, the vapor state. A vapor is a substance at or near its condensation point. It may be wet, dry, or slightly superheated. One hundred percent dryness is an exactly definable condition which is only transiently encountered in practice. Vapor behavior deviates so widely from the ideal gas laws that calculation requires the use of tables and graphs that give the experimentally determined properties of the fluid.

Power and refrigeration plants. A steam power plant operates on a vapor cycle where steam is generated by boiling water at high pressure, expanding it in a prime mover, exhausting it to a condenser, where it is reduced to the liquid state at low pressure, and then returning the water by a pump to the boiler (**Fig. 1**).

In the customary vapor-compression refrigeration plant, the process is essentially reversed with the refrigerant evaporating at low temperature and pressure, being compressed to high pressure, condensed at elevated temperature, and returned as liquid refrigerant through an expansion valve to the evaporating coil (**Fig. 2**).

The Carnot cycle, between any two temperatures, gives the limit for the efficiency of the conversion of heat into work (**Fig. 3**). This efficiency is independent of the properties of the working fluid. Although the thermal efficiency is independent of the

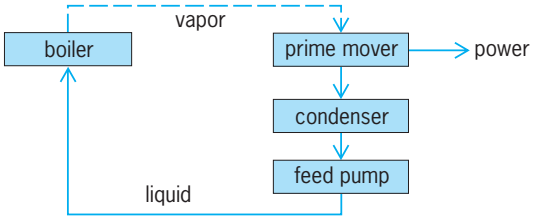


Fig. 1. Rudimentary steam power plant flow diagram.

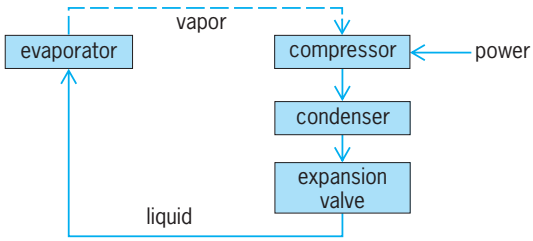


Fig. 2. Rudimentary vapor-compression refrigeration plant flow diagram.

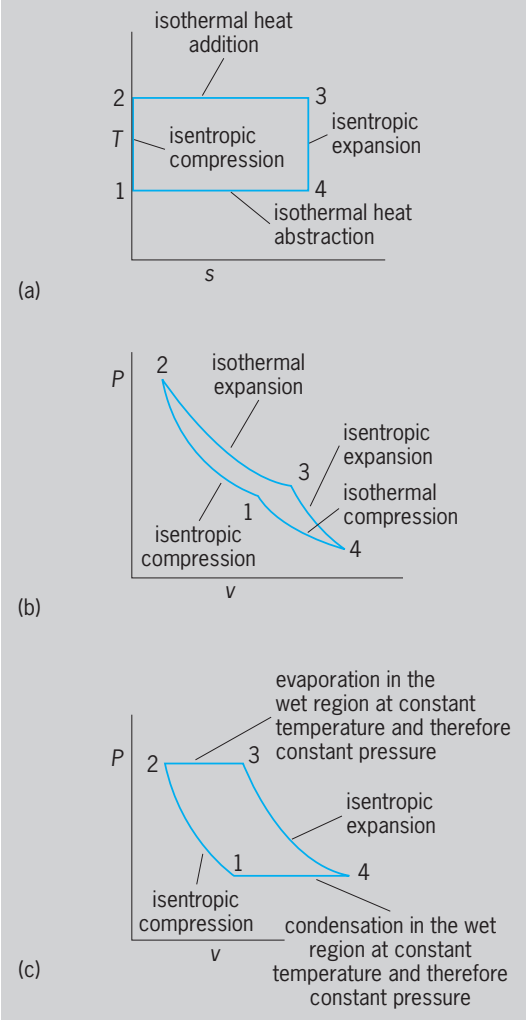


Fig. 3. Carnot cycle. (*a*) Temperature-entropy. (*b*) Pressure-volume for fixed gas. (*c*) Pressure-volume for vapor.

properties of the fluid, the mean effective pressure (and consequent physical dimensions of the engine) will be vitally influenced by choice of fluid (compare Fig. 3*b* with *c*). The Carnot cycle is not realistic for the evaluation of steam power plant performance because the cycle precludes the use of superheat and calls for the isentropic compression of vapor. It is useful, however, for specifying the limiting efficiency that a real cycle might approach. It is so used in judging performance of vapor-cycle heat engines and heat pumps. *See* CARNOT CYCLE.

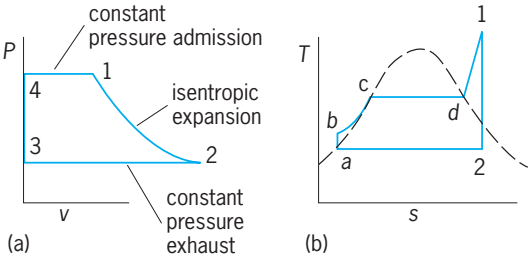


Fig. 4. Rankine cycle for heat-engine plant. (*a*) Pressure-volume diagram of prime mover. (*b*) Temperature-entropy diagram for the power plant.

The Rankine cycle is more realistic in describing the ideal performance of steam power plants and vapor-compression refrigeration systems.

Vapor steam plant. In the case of the steam power plant (Fig. 1), the Rankine cycle (**Fig. 4**) has two constant pressure phases joined by a reversible adiabatic (isentropic) phase 1–2. From the properties of the fluid, the work of the prime mover, ΔW_{PM}, is most conveniently evaluated as in Eq. (1), where h

$$\Delta W_{PM} = h_1 - h_2 \tag{1}$$

is the enthalpy, Btu/lb. The feed pump uses some of this work, ΔW_{FP}, to return the water from the condenser to the boiler so that the net output of the cycle is $\Delta W_{PM} - \Delta W_{FP}$. This net output can be related to the heat that must be added to produce steam by consideration of the *T-s* diagram (Fig. 4*b*). The area under line *bcd*–1 is the heat supplied in the boiler (heat source); phase 1–2 is the isentropic expansion in the prime mover; the area under line 2–*a* is the heat rejected to the condenser (heat sink); phase *a*–*b* is the isentropic compression of the liquid in the feed pump. Thus, the thermal efficiency is given by Eq. (2).

$$\frac{\text{Work done}}{\text{Heat added}} = \frac{\Delta W_{pm} - \Delta W_{FP}}{h_1 - h_a - \Delta W_{FP}} \tag{2}$$

There are many variables which influence the performance of the Rankine cycle. For steam the thermal efficiency is a function of pressure, temperature, and vacuum (**Fig. 5**). High pressure, high superheat, and high vacuum lead to high efficiency. *See* RANKINE CYCLE.

Vapor refrigeration plant. The Rankine cycle can be used to evaluate the performance of the vapor-compression system of refrigeration (Fig. 2). A counterclockwise path is followed on the *P-v* and *T-s* cycle diagrams (**Fig. 6**). The refrigerant enters the compressor as low-temperature, low-pressure vapor (4–1); isentropic compression follows (1–2), and then high-pressure delivery (2–3). The work to drive the compressor is $h_2 - h_1$. The machine cooling coefficient of performance *cp*, which is essentially the reciprocal of thermal efficiency, is given by Eq. (3); for a warming machine, it is given by Eq. (4).

$$cp = \frac{\text{refrigeration}}{\text{work done}} = \frac{h_1 - h_d}{h_2 - h_1} \tag{3}$$

$$cp = \frac{\text{heat delivered}}{\text{work done}} = \frac{h_2 - h_c}{h_2 - h_1} \tag{4}$$

The difference $h_1 - h_d$ is heat removed in the refrigerating coils, the area under phase *d*–1 on the *T-s* diagram. The difference $h_2 - h_c$ is the heat delivered to the condensing coils, the area under phase (2*abc*). Because the flow is throttled through the expansion valve, the enthalpy is constant, $h_c = h_d$. Some ideal performance values of a vapor-compression refrigeration system are plotted in **Fig. 7**. For further details on heat-pump vapor cycles *see* HEAT PUMP.

The remainder of this article is concerned with

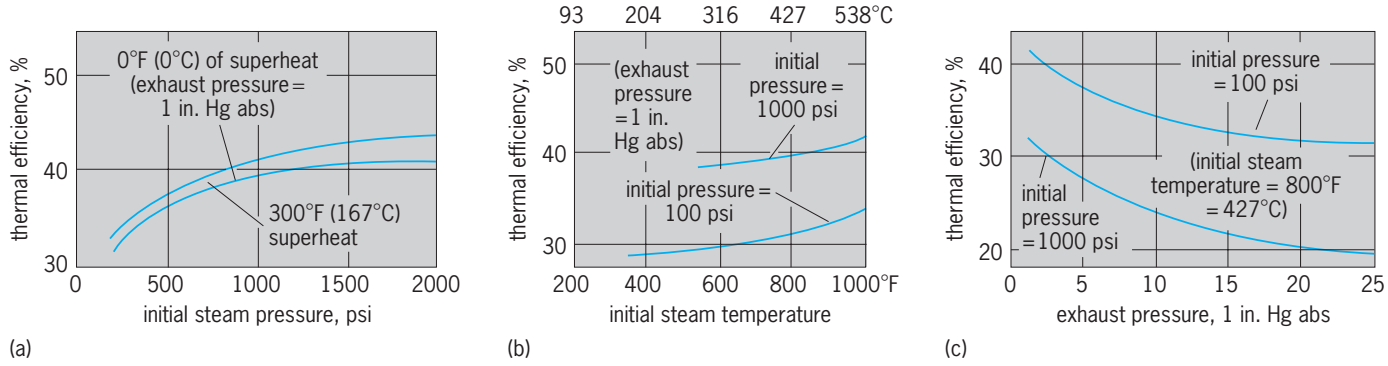


Fig. 5. Thermal efficiency of ideal Rankine steam cycle. (*a*) Steam pressure. (*b*) Steam temperature. (*c*) Vacuum. 1 psi = 6.9 kPa; 1 in. Hg abs = 3.4 kPa.

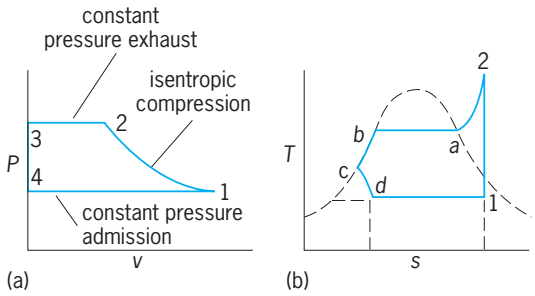


Fig. 6. Rankine cycle for heat-pump plant. (*a*) Pressure-volume diagram of compressor. (*b*) Temperature-entropy diagram for a vapor-compression refrigeration plant.

the vapor cycle as it is applied to power-generation purposes.

Regenerative heat cycle. The regenerative cycle is a modification of the simple Rankine cycle. Feed water is heated by extracted steam (**Fig. 8**). As a result, less heat must be added in the boiler to evaporate a pound of steam and in turn to deliver a kilowatt hour of work output from the associated steam engine.

A cycle with a single stage of regenerative heating can be viewed as two Rankine cycles superimposed on one another (Fig. 8*b*). In one the exhaust pressure and consequent temperature are substantially higher than in the other (point *f* versus point 2). The heat of condensation represented by the area under *c*–*f* can be used to raise the feed temperature from T_b to T_c. The area under phase *b*-*c* is smaller than under phase *c*–*f* so a fraction of a pound of steam is needed to raise 1 lb of water to the common temperature level T_c.

The principle of regeneration can be extended to multiple-stage heating with different final feed temperatures and in the limit reaching the boiler saturation temperature T_c with an infinite number of heating stages. Some consequences of the process are reflected in the data of **Fig. 9**. The gain in thermal efficiency is the consequence of reducing the quantity of heat rejected to sink 2–*a* in Fig. 8. The weight of steam flow to the prime mover for the production of a kilowatt-hour is larger than with the simple Rankine cycle (Fig. 4). But the heat required to make a pound of steam is so much less that there is an overall thermodynamic gain per kilowatt-hour.

The weight flow of steam, from the prime mover to the condenser, is less than with the simple Rankine cycle. It is this reduction in heat rejected to the thermodynamic sink that raises the overall thermal efficiency of the regenerative above the nonregenerative cycle. Modern steampower practice uses the steam turbine for up to ten stages of regenerative heating.

Reheat cycle. The resuperheat or reheat cycle is another improvement in vapor cycles favored in current central station practice. Steam expanding

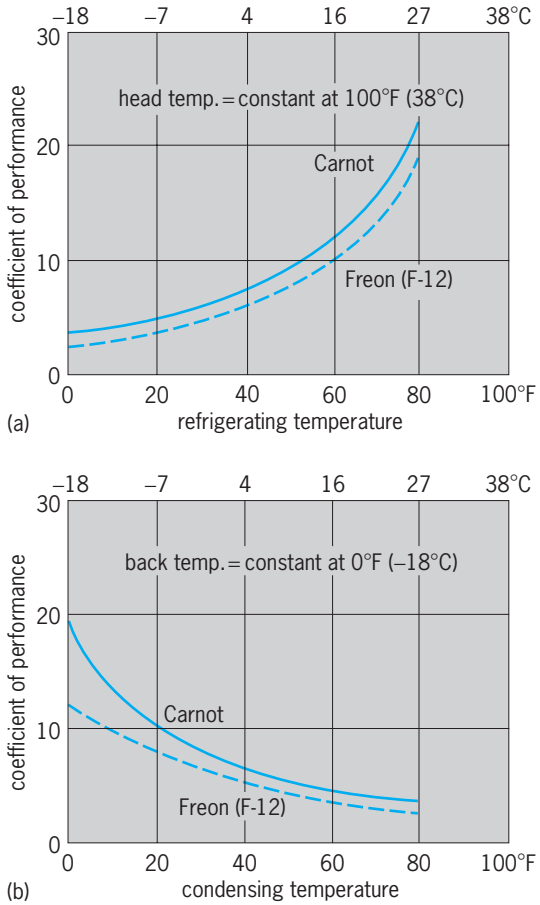


Fig. 7. Coefficient of performance of ideal Rankine and Carnot cycles as influenced by (*a*) refrigerating temperature and (*b*) condensing temperature.

isentropically (Fig. 4*b*) grows wetter with consequent increased erosion of machinery parts and loss in mechanism efficiency: Superheat tends to correct these weaknesses but metallurgical limitations fix the maximum allowable steam temperature. Reheating the steam after a partial expansion (**Fig. 10***a*) gives a practical correction. The reheating can be carried out at various pressure and temperature levels and in multiple stages. Thermal efficiency is improved over the simple Rankine cycle (Fig. 10*b*). Current practice uses a single reheat stage with temperatures approximately equal to primary steam temperature. Supercritical pressure plants favor two reheating stages.

Binary vapor cycle. Comparison of the Rankine (Fig. 4*b*), the reheat (Fig. 10*a*), and Carnot (Fig. 3*a*) cycles shows that there are considerable thermodynamic losses in the first two by failure to approach the rectangular *T-s* configuration of the Carnot cycle. The binary vapor cycle uses two fluids with totally different vapor pressures, such as mercury and water (**Fig. 11**). If a Rankine cycle using mercury is superimposed on that using steam, it is possible to operate the mercury condenser as the steam boiler, transferring the heat from the one fluid to the other. Because of the differences in the latent heats and specific heats of the two fluids, several pounds of mercury must be circulated to make 1 lb of steam. However, the combined

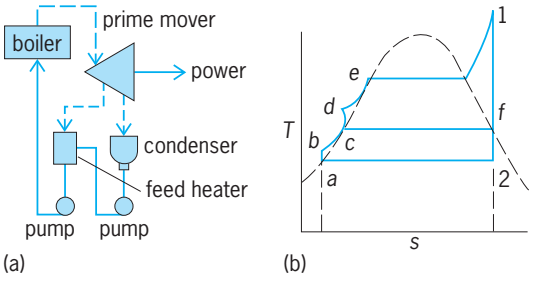


Fig. 8. Regenerative cycle. (*a*) Rudimentary steam power plant flow diagram with single-stage feed heating. (*b*) Temperature-entropy diagram.

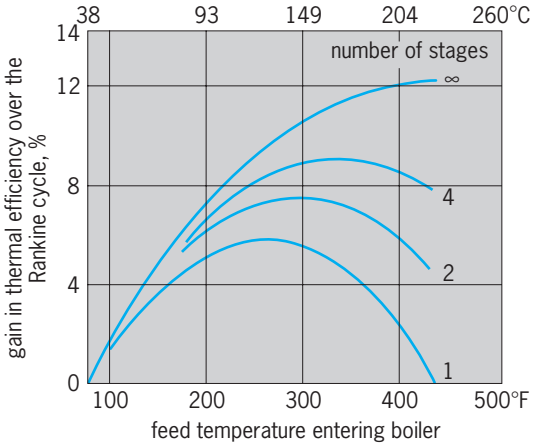


Fig. 9. Gain in thermal efficiency by use of regenerative instead of Rankine cycle; steam conditions, 400 psi (2.76 MPa) and 700°F (371°C); exhaust pressure, 1 in. Hg abs (3.4 kPa).

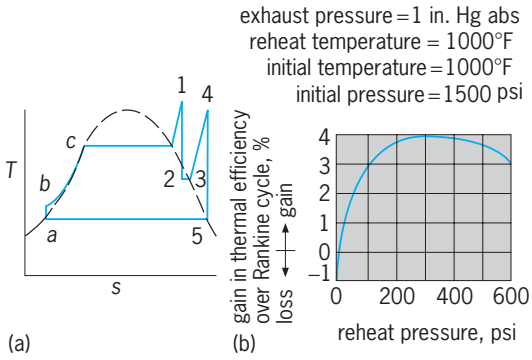


Fig. 10. Resuperheat or reheat cycle. (*a*) Temperature-entropy diagram. (*b*) Gain in thermal efficiency as function of reheat pressure; primary pressure, primary temperature, and reheat temperature constant at 1500 psi (10.3 MPa), 1000°F (538°C), and 1000°F (538°C), respectively. 1 in. Hg abs = 3.4 kPa; 1 psi = 6.9 kPa.

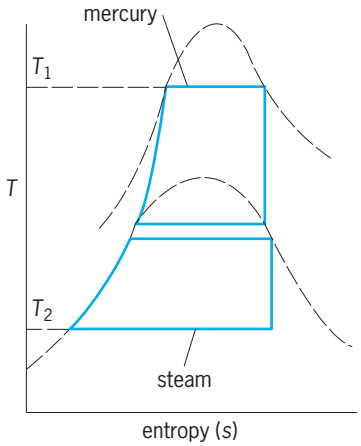


Fig. 11. Binary vapor cycle using mercury and water, temperature-entropy diagram.

T-s diagram (Fig. 11) approximates the rectangular specification with appreciable gain in thermal efficiency. For many years the binary vapor cycle was thought to have practical as well as theoretical advantages, but its economic difficulties have led to its eclipse by, and abandonment in favor of, the high-pressure, regenerative, reheat, steam cycles. *See* THERMODYNAMIC CYCLE.

Theodore Baumeister

Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; V. M. Faires, *Thermodynamics*, 6th ed., 1978; W. L. Haberman and J. E. John, *Engineering Thermodynamics*, 2d ed., 1989; K. A. Rolle, *Thermodynamics and Heat Power*, 4th ed., 1994; J. R. Simonson, *Engineering Heat Transfer*, 2d ed., 1989.

Vapor lamp

A source of radiant energy excited by a supply of electricity which creates a current of ionized gas between electrodes in an enclosure that contains the arc while permitting transmission of the radiant energy. Gaseous-discharge lamps or vapor lamps are given various names relating to the element

Approximate efficiencies of popular lamp types

Lamp	Nominal efficiency, lumens per watt
Mercury vapor	50
Fluorescent	70
Metal-halide	90
High-pressure sodium	110
Incandescent	20

responsible for the majority of the radiation (mercury, sodium metal-halide, xenon), to the physical attribute of the lamp (short-arc, high-pressure), or, in the case of fluorescent lamps, to the way a phosphor on the bulb wall fluoresces as a result of the lamp's low-pressure mercury-vapor excitation. *See* ARC DISCHARGE; ELECTRICAL CONDUCTION IN GASES.

Gaseous-discharge lamps are broadly and increasingly used throughout the world because the conversion of electric energy to radiant energy in a gaseous discharge provides radiation in narrow bands within the range of visible light in which the rods and cones of the eye are most sensitive. These light sources have high efficiency in conversion of electricity to light (see **table**). The popularity of high-pressure sodium lamps that emit yellow-white light for roadway lighting is due to high efficiency. This lamp's principal radiation is near the peak sensitivity of the eye, while radiation at other wavelengths is produced through a spectrum-line broadening due to high pressure.

The gaseous-discharge lamp requires a high voltage to start, but as the electric current increases, the resistance decreases, and some means must be available to limit the current and avoid lamp failure. Some gaseous-discharge lamps use a tungsten filament within the lamp envelope, which gets hot and limits the current. More commonly, a magnetic structure called a ballast (transformer) is used external to the lamp and limits the lamp current. *See* ILLUMINATION; TRANSFORMER.

For discussions of common types of vapor lamps *see* FLUORESCENT LAMP; MERCURY-VAPOR LAMP; METAL HALIDE LAMP; NEON GLOW LAMP; SODIUM-VAPOR LAMP; SUNLAMP; ULTRAVIOLET LAMP.

T. F. Neubecker

Vapor lock

Interruption of fuel flow to an engine due to blockage of passages in the fuel system by fuel vapor.

To promote easy starting, all gasolines contain volatile constituents that under some conditions, such as high ambient temperature, tend to produce more vapor than the fuel-system vents can handle. The action of an engine-mounted fuel pump, in decreasing the pressure at its inlet, tends to vaporize the fuel. If the vapor forms faster than the pump can draw it from the fuel line, the flow of fuel to a carburetor is effectively stopped and the engine stalls. Vapor lock is much less likely on a fuel-injected engine with an electric pump in the fuel tank. However, an engine with port fuel injection may experience vapor lock if the injector or fuel overheats, or if the Reid vapor pressure of the fuel is too high. *See* CARBURETOR; INTERNAL COMBUSTION ENGINE; VAPOR PRESSURE.

The tendency to form vapor in a given fuel system has been correlated with the Reid vapor pressure and ASTM distillation curve of the gasoline. With a given fuel, vapor formation can be minimized by reducing heat in the fuel system, increasing fuel-system pressure, and eliminating sudden changes in cross section or direction of fuel lines. *See* AIRCRAFT FUEL; FUEL SYSTEM; GASOLINE.

Donald L. Anglin

Bibliography. W. H. Crouse and D. L. Anglin, *Automotive Engines*, 8th ed., 1994.

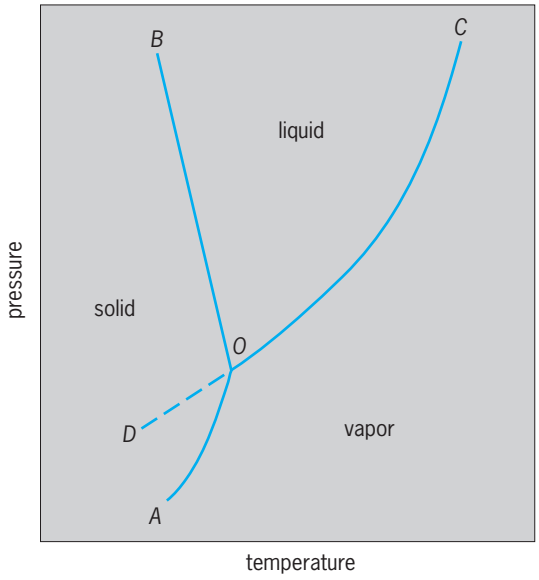
Vapor pressure

The saturation pressures exerted by vapors which are in equilibrium with their liquid or solid forms. One of the most important physical properties of a liquid, the vapor pressure, enters into many thermodynamic calculations and underlies several methods for the determination of the molecular weights of substances dissolved in liquids. For a discussion of the vapor pressure relationships of solids *see* SUBLIMATION; MOLECULAR WEIGHT; SOLUTION.

If a liquid is introduced into an evacuated vessel at a given temperature, some of the liquid will vaporize, and the pressure of the vapor will attain a maximum value which is termed the vapor pressure of the liquid at that temperature. Although the quantity of liquid remaining does not diminish thereafter, the process of evaporation does not cease. A dynamic equilibrium is established, in which molecules escape from the liquid phase and return from the vapor phase at equal rates. *See* EVAPORATION.

It is important to make a distinction between the vapor pressure of a liquid, as described above, and the pressure of a vapor. The vapor pressure of a pure liquid is a unique and characteristic property of the liquid and depends only upon the temperature. A gas or vapor may, on the other hand, exert any pressure within reason, depending upon the volume to which it is confined, provided it is not in contact with its liquid phase.

Liquid-vapor equilibrium. The relationship between the vapor pressure of a liquid and the temperature is indicated by a phase diagram. The **illustration** shows the phase diagram for water, the line *OC* being the vapor pressure line for liquid water. Line *AO* is the vapor pressure line (sublimation pressure curve) for ice, and line *BO* is the liquid-solid equilibrium line. Point O is called the triple point and is the unique pressure and temperature at which a pure solid, its liquid, and its vapor can coexist in equilibrium under the pressure of the vapor alone. The triple point of water is not the familiar melting point (0°C or 32°F) but rather 0.01°C (32.018°F). The distinction is that the total pressure at the triple point is 4.58 mm (611 pascals), the common vapor pressure of solid and liquid water, whereas the



Phase diagram of the water system.

total pressure at the melting point is ordinarily 1 atm (100 kilopascals). Point *C* is called the critical point and is the point above which there is no distinction between the liquid and gaseous phases. It is not possible to liquefy a gas at temperatures above the critical temperature, regardless of the applied pressure. Necessarily, the surface tension and the latent heat of vaporization become zero at the critical point. For ordinary liquids, the vapor pressure at the critical point is usually about 50 atm (5000 kPa). There is no evidence for a critical point on the solid-liquid equilibrium line, and *B* is meant to indicate a direction, rather than a point; other phases may appear above *B*, of course, and alter the direction of the line. It is possible to undercool a liquid below its triple point if crystallization nuclei are absent. The vapor pressure of the undercooled liquid is given by the broken line *DO* and lies above the equilibrium vapor pressure of the solid. The undercooled liquid is therefore metastable, because the system tends to assume its lowest vapor pressure at equilibrium. *See* CRITICAL PHENOMENA; TRIPLE POINT.

Quantitative relations. For most liquids the relationship between the vapor pressure and temperature can be expressed by an equation having the form of (1), where a, b, and c are constants. The simpler equation (2), where m and n are constants, is often

$$\log P = a + b \log T + \frac{c}{T} \tag{1}$$

$$\log P = \frac{m}{T} + n \tag{2}$$

adequate. The change in the vapor pressure of a liquid with temperature may be expressed by the Clausius-Clapeyron equation (3), where L_v is the

$$\frac{dP}{dT} = \frac{L_v}{T\Delta V} \tag{3}$$

molar latent heat of vaporization and ΔV is the difference in molar volumes of the vapor and liquid at the temperature T. Because the molar volume of the liquid is negligible by comparison with that of the vapor except near the critical point, the Clausius-Clapeyron equation may be written as (4), where V

$$\frac{dP}{dT} = \frac{L_v}{TV} \tag{4}$$

is the molar volume of the gas. If the gas follows the ideal gas law ($PV = RT$), the equation may be further written as (5), where R is the gas constant. For

$$\frac{1}{P}\frac{dP}{dT} = \frac{L_v}{RT^2} \tag{5}$$

moderate ranges of temperature, the latent heat of vaporization is constant, and the equation may be rearranged and integrated to give the useful form of (6).

$$\log_{10} \frac{P_2}{P_1} = \frac{L_v}{2.3R}\left(\frac{1}{T_1} - \frac{1}{T_2}\right) \tag{6}$$

From the latent heat of vaporization and the vapor pressure at one temperature, the vapor pressures at other temperatures may thus be calculated.

The vapor pressure of a liquid is lowered when a substance is dissolved in it. At low solute concentrations in nonideal solutions, or at all solute concentrations in ideal solutions, the partial vapor pressure of the liquid in a solution is proportional to the mole fraction of the solvent, Eq. (7), where X_1 is the sol-

$$P_1 = X_1 P_1^0 \tag{7}$$

vent mole fraction and P_1^0 is the vapor pressure of the pure solvent. Raoult's law, as this important relationship is known, may also be expressed in the form of Eq. (8), where X_2 is the mole fraction of the solute.

$$\frac{P_1^0 - P_1}{P_1^0} = X_2 \tag{8}$$

The lowering of the vapor pressure of the solvent is proportional to the mole fraction of the solute. This relation provides the basis for the determination of the molecular weights of dissolved substances. *See* CONCENTRATION SCALES.

Vapor pressure measurement. Because the vapor pressures of liquids range widely, a number of methods have been devised for their measurement. The static method consists of introducing an excess of the liquid into an evacuated system at a given temperature, and measuring the vapor pressure with an attached mercury manometer. It is useful for vapor pressures ranging from a few millimeters of mercury up to several atmospheres. The dynamic method consists of introducing the liquid into a system in which the pressure may be varied, and noting the temperature at which the liquid boils at a given pressure. It is useful for moderately high vapor pressures. The transpiration method consists of bubbling a known volume of an inert gas at a definite pressure through the liquid. The quantity of liquid transported into the carrier gas is determined, and the vapor pressure of the liquid is calculated from

relation (9), where P_C is the carrier gas pressure

$$P_L = P_C \frac{n_L}{n_L + n_C} \tag{9}$$

and n_L and n_C are the numbers of moles of liquid vaporized and carrier gas, respectively. For the measurement of very low vapor pressures, the Knudsen effusion method may be used. This method is based upon the measurement of the mass of vapor which escapes through a very small hole in a vessel containing a liquid in equilibrium with its vapor. If the diameter of the hole is small compared with the mean free path of the gas molecules, the latter suffer no collisions in their passage through the hole. Equation (10) relates the mass of gas which passes

$$m = P\sqrt{\frac{M}{2\pi RT}} \tag{10}$$

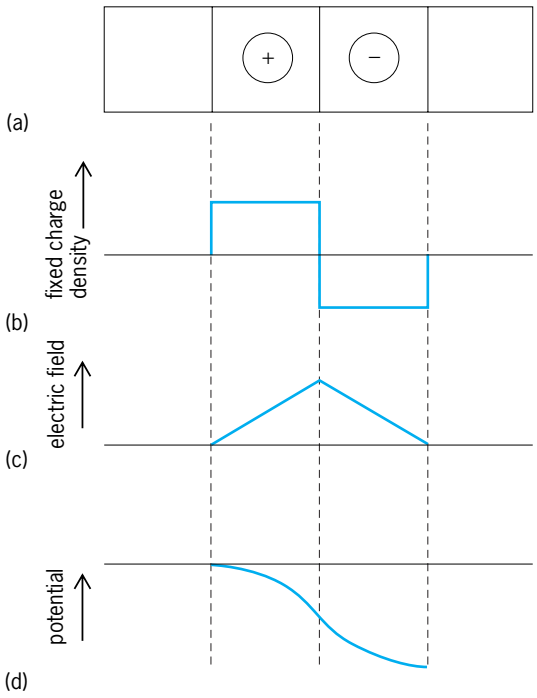
per square centimeter per second through the hole and the vapor pressure, where m is the mass of gas per square centimeter per second and M is the molecular weight. The Knudsen method has been used with radioactive isotopes or with mass spectrometers to determine very low vapor pressures. *See* PHASE EQUILIBRIUM. Norman H. Nachtrieb

Bibliography. P. W. Atkins, *Physical Chemistry*, 6th ed., 1997; G. M. Barrow, *Physical Chemistry*, 6th ed., 1996.

Varactor

A solid-state device which has a capacitance that varies with the voltage applied across it. The name varactor is a contraction of the words variable and reactor. Typically the device consists of a reverse-biased *pn* junction that has been doped to maximize the change in capacitive reactance for a given change in the applied bias voltage. The device has two primary applications: frequency-tuning of radio-frequency circuits including frequency-modulation (FM) transmitters and solid-state receivers, and nonlinear frequency conversion in parametric oscillators and amplifiers. *See* CAPACITANCE; FREQUENCY MODULATOR; PARAMETRIC AMPLIFIER; REACTANCE.

A *pn* junction in reverse bias has two adjacent microscopic space-charge or depletion regions which function like the plates of a capacitor (see **illus.**). These depletion regions get larger as the applied reverse-bias voltage is increased; however, the increase in the width of the depletion region is not linear with bias voltage, but instead is sublinear, the exact nature of the relationship depending upon the doping profile in the *pn* junction. For example, in a *pn* junction with constant doping density, the depletion region width varies as the square root of the applied reverse-bias voltage. Because the capacitance of the device is proportional to the width of the depletion region, the nonlinear relationship between bias voltage and depletion width results in a nonlinear voltage-capacitance relationship as well. The *pn* junction doping profile is adjusted by the device designer to obtain the desired capacitive nonlinearity.



Depletion region of a *pn* junction used in a varactor diode. (*a*) Fixed charge. (*b*) Density of fixed charge stored capacitively in the depletion layer. (*c*) Built-in electric field which results from formation of the depletion region. (*d*) Electric potential.

The frequency response of the varactor is governed by the relationship between the series linear resistance of the diode and its nonlinear capacitance. The highest frequency for which the device will function properly is that at which the capacitive reactance (the reciprocal of the product of nonlinear capacitance and frequency) is equal to the series resistance of the device. Thus, designing a varactor for maximum frequency response involves choosing a doping density high enough for a small series resistance but low enough so that the capacitance of the device is small.

Varactors, as well as other solid-state devices, possess the advantage that they are compact and robust, permitting their use in hostile environments, as well as improving the reliability of the circuits in which they are employed. *See* JUNCTION DIODE; MICROWAVE SOLID-STATE DEVICES; SEMICONDUCTOR; SEMICONDUCTOR DIODE. David R. Andersen

Bibliography. D. Christiansen (ed.), *Electronics Engineers' Handbook*, 4th ed., 1996; D. H. Navon, *Semiconductor Microdevices and Materials*, 1986; L. S. Nergaard and M. Glicksman (eds.), *Microwave Solid State Engineering*, 1964; E. S. Yang, *Microelectronic Devices*, 1988.

Variable (mathematics)

A term that became part of mathematical language during the development of analytic geometry (no equivalent word is found in the writings of Archimedes and other Greek mathematicians), and which is frequently not defined in modern works,

apparently in the belief that the meaning ordinarily attached to the word is sufficiently precise. When a definition is given, it is usually to the effect that a symbol x is a variable if it may denote any member of a set S of objects. A variable is discrete or continuous according as its range (the set S) is discrete (for example, a subset of the natural numbers) or continuous (for example, all real numbers between two real numbers), respectively. Such a working definition usually suffices for the needs of mathematics. *See* ANALYTIC GEOMETRY; NUMBER THEORY; PARAMETRIC EQUATION.

Leonard Blumenthal

Variable star

A star with a detectable change in brightness that is often accompanied by other physical changes. Every star varies in brightness sometime during its life, usually in the early stages (while it is forming) or in the late stages (close to its death). Therefore, variability provides important clues about the evolution and nature of stars. Depending upon the type of star, variability in brightness can provide information about its size, radius, mass, temperature, luminosity, internal and external structure, composition, and distance from the Earth.

Hundreds of thousands of stars are known to vary in brightness. Some variable stars are very bright and can be observed with the unaided eye, while others are observable only with telescopes. The variation from maximum to minimum brightness (the range) can be anywhere from thousandths of a magnitude to many full magnitudes, and the period can be anywhere from a fraction of a second to several years. The change in brightness may be caused by intrinsic physical changes in the star or stellar system such as pulsations or eruptions; extrinsic changes, such as the eclipse of the star by one or more stars; or rotation of a star with bright or dark spots on its surface. *See* MAGNITUDE (ASTRONOMY).

Observation. Most variable stars need to be monitored in a systematic way over years and decades in order to determine their period, long-term behavior, and unusual and brightness variation changes. The simplest method of determining the brightness of variable stars is visually through a telescope or binoculars. Groups of observers worldwide, such as the American Association of Variable Star Observers (AAVSO), carry out systematic visual brightness determinations for thousands of variable stars. Group members, mostly amateur astronomers, observe with their own binoculars and telescopes. They use special star maps (finding charts) that identify the variable star and stars of constant and known brightness (comparison stars) used in estimating the brightness of the variable. Observational data provided by these dedicated observers have been used extensively by professional astronomers to analyze the behavior of the changes in brightness of variable stars, to schedule and correlate observations made with large telescopes and special instruments (both on the Earth and aboard satellites orbiting the Earth), and to make theoretical variable-star models with the aid of powerful computers in order to understand better the physical structure of these stars.

Other techniques for systematically monitoring variable stars, used by both professionals and increasing numbers of amateurs, include photographic and photoelectric photometry and charge-coupled devices (CCDs). Photographic plates provide a lasting record of the brightness of stars. Photoelectric photometry uses photomultipliers to amplify and measure the intensity of the light from a star very precisely. It is particularly useful for observing stars that have a small range of variation, from thousandths to tenths of a magnitude. CCDs are semiconductor imaging devices consisting of a two-dimensional array of photodetectors and a readout amplifier. They convert optical images into electrical signals, and are particularly useful in providing images of faint objects not seen through the eyepiece of a telescope. *See* ASTRONOMICAL IMAGING; CHARGE-COUPLED DEVICES; PHOTOMETRY; PHOTOMULTIPLIER.

Certification and nomenclature. Once a star is discovered to be changing in brightness, and its position in the sky and the kind (type) and the amount (range) of its variability are determined, it is ready to be certified and named as a variable star. The naming of a variable star is based on a system developed in the mid-1800s by F. W. A. Argelander. The first variable star discovered in a constellation is given the letter R, followed by the genitive form of the Latin name of the constellation. For example, R Leonis is the variable star R in the constellation of Leo. When other variables are discovered in that constellation, they receive the letters S, T, and so on to Z. Subsequent variables are named RR, RS to RZ, then SS to SZ, and so on to ZZ, after which follow AA, AB to AZ, and so on to QZ. (The letter J is never used in this system.) The next variable after QZ, the 334th variable in a constellation, is named V335. (For example, the variable V335 Sagittarii is the 335th variable in Sagittarius.) Stars that have been assigned Greek letters, such as o (omicron) Ceti, or small roman letters, such as g Herculis, prior to the start of this system continue to keep those names.

Until World War II, the certification and naming of a variable was done in Germany. The International Astronomical Union continued the certification and naming until 1952. This task has been carried out by the compilers of the *General Catalogue of Variable Stars* in Russia ever since. The fourth edition of this catalog and name lists that supplement it comprise 38,622 known variables designated through 2003. An additional 14,810 stars listed in the *New Catalogue of Suspected Variables* (1982) are in need of confirmation of their variability.

Not all constellations have the same number of variables, because the distribution of variables is not uniform in the Milky Way Galaxy. While the concentration is very high in the disk, particularly toward the center of the Galaxy (as in the case of the constellation Sagittarius, with 4374 variable stars listed),

the corona of the Milky Way Galaxy has the lowest number (as in the case of the constellation Caelum, which has only nine variables listed). *See* MILKY WAY GALAXY.

With the advent of CCDs, thousands of new faint variable stars are being discovered, and additional thousands of stars which vary only by millimagnitudes are being discovered as well with sensitive photoelectric photometers aboard space satellites. Yet, there may still be thousands of variable stars undiscovered, too faint, with too small variability, or too distant to be detected with present instrumention. There are several ongoing surveys of the sky, such as the All-Sky Automated Survey (ASAS), that are finding new variables every day. A popular activity for amateur astronomers is data mining the new surveys to search for unique objects for further study. *See* DATA MINING.

Classification. The light curve and the spectrum of a variable star form the basis for determining its classification. A light curve (**Fig. 1**) is the graph of the observed brightness (usually in magnitudes) of the star plotted against time (usually in Julian days). The important parameters are the shape of the light curve, the brightness range, and the regularity and the period of the brightness variation. *See* LIGHT CURVES.

The spectrum of a star helps to measure its temperature, luminosity, chemical composition, age (population), and mass. The spectra of variable stars normally range from the O or B (very hot/blue), to the M, C, and R (cool/red) spectral classes. Brown dwarfs (L and T spectral classes) also exhibit variability. Some lines in the spectra help to identify the nature of the variable. For example, metallic oxides and carbon lines, accompanied by bright lines of hydrogen, are usually seen in cool, red, long-period variables. *See* ASTRONOMICAL SPECTROSCOPY; BROWN DWARF; SPECTRAL TYPE.

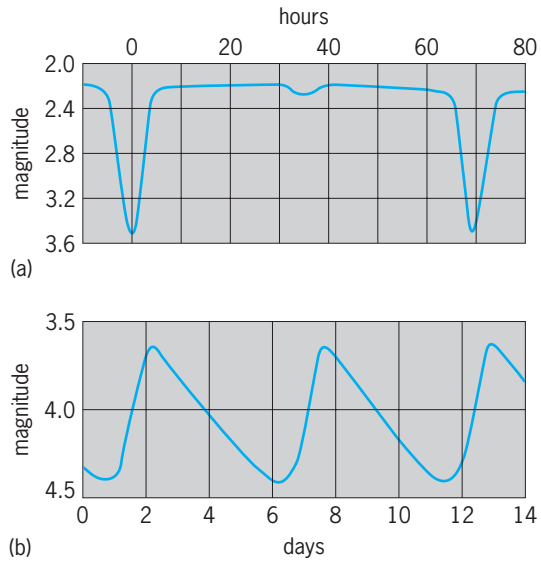


Fig. 1. Comparison of eclipsing variable stars and intrinsic variable stars. (*a*) An eclipsing variable star (Algol), with a period of 2.87 days. (*b*) A typical intrinsic variable star (Delta Cephei), with a period of 5.37 days.

Variable stars can be divided into two major types: extrinsic and intrinsic variables (Fig. 1).

Extrinsic variables. The extrinsic variables are those stars in which the variability in brightness occurs because of the occultation of one star by another object, or by the rotation of a spotted star.

In the eclipsing binary type of star, the change in the brightness of the system is the result of a geometric phenomenon. The primary eclipse occurs when the brighter star is eclipsed by the fainter, and the secondary eclipse occurs when the fainter star is eclipsed by the brighter. Information on the size, shape, and the distance to the stars from the Earth may be obtained from the eclipse phenomenon of these stars. The prototype of eclipsing binaries is Algol (β Persei), a bright star, observable with an unaided eye, that changes its brightness from 2.1 to 3.4 magnitudes as it is eclipsed every 69 hours (Fig. 1*a*). *See* BINARY STAR; ECLIPSING VARIABLE STARS.

A second type of eclipsing object is the extrasolar planetary transit. Here, a planet transits across the disk of its parent star, causing a dip in brightness similar to an eclipsing binary star. The dip is usually very small, as planets are much smaller than the star they orbit, but is detectable by careful ground-based photometry. *See* EXTRASOLAR PLANETS; TRANSIT (ASTRONOMY).

In a rotating variable, the variability is caused by rapid rotation of a star, often a member of a binary system, that has dark or bright spots on its surface, similar to sunspots. *See* BINARY STAR; ECLIPSING VARIABLE STARS.

Intrinsic variables. Intrinsic variables are those stars in which the variability of brightness occurs because of physical change in or on the star itself. These stars are divided into two classes: pulsating and eruptive (see **table**).

Pulsating variables. Periodic pulsation (contraction and expansion) of the star and its outer layer results in the variation in brightness, as well as variations in the star's temperature, spectrum, and radius, for pulsating variables, Cepheids are rare, highly luminous (supergiant), yellow pulsating stars. They vary with periods from 1 to 100 days, and have a range of variation from 0.1 to 2 magnitudes. The prototype of the class is δ Cephei, a star observable with an unaided eye which varies in brightness from magnitude 3.7 to 4.4 every 5.4 days (Fig. 1*b*). Polaris, the North Star, is also a cepheid variable with a smaller range of variation from magnitude 2.5 to 2.6 every 4 days. *See* POLARIS.

Cepheids show an important correlation between their period of variation and their relative brightness, in that those stars with longer periods are also brighter. This period-luminosity relationship was first noticed in 1912 by H. Leavitt while studying the cepheid variables in the Small Magellanic Cloud (**Fig. 2**). H. Shapley extended this relationship to cepheids in the Galaxy, and converted the relationship to period and intrinsic luminosity of these stars. In 1950, W. Baade determined the zero point of this

Representative types of intrinsic variables

Type	Range of period, days	Amplitude, magnitude	Spectra and luminosity	Example	Number known
Pulsating					
RR Lyrae	<1	<2	A to F (blue), giants	RR Lyrae	7382
Cepheids (types I and II)	1–70	0.1–2	F to G (yellow), supergiants	δ Cephei	1028
Long-period					
Mira	80–1000	2.5–6+	M (red), giants	o Ceti (Mira)	7390
Semiregular	30–1000	1–2	M (red), giants and supergiants	Z Ursae Majoris	5479
RV Tauri	30–150	Up to 3	G to K (yellow and red), supergiants	RV Tauri	134
Irregular	Irregular	Up to several magnitudes	All types		5884
Others					707
Eruptive					
Supernovae	?	15 or more		CM Tauri (Crab Nebula)	7
Novae	Centuries?	7–16	O to A, subdwarfs	GK Persei	388
Recurrent novae	20 years?	7–9		RS Ophiuchi	8
Dwarf novae	10–600	2–6	A to F, subdwarfs	SS Cygni, Z Camelopardalis	420
Flare	?	Up to 6	M, dwarfs	UV Ceti	1780
Nebular	Rapid and erratic	Up to a few magnitudes	B to M, main sequence and subgiants	T Tauri	1698
R Coronae Borealis		1–9	F to K, supergiants	R Coronae Borealis	42
Others					380

relationship, with his work on these stars in the Andromeda Galaxy. *See* ANDROMEDA GALAXY; MAGELLANIC CLOUDS.

Thus, by observing the apparent brightness change and determining the period of this variation, use of the period-luminosity relationship leads to the absolute magnitude of the star. Once the absolute magnitude of a star is known, the distance to it can be determined from the equation below,

$$M = m + 5 - 5 \log D$$

where M is the absolute magnitude, m is the apparent magnitude, and D is the distance in parsecs. Due to this important correlation, cepheids have been used to measure distances to nearby galaxies and in the determination of the distance scale of the universe.

A similar group of variable stars, W Virginis stars, are found in globular star clusters and the corona of the Milky Way Galaxy. These stars are bluer (and thus hotter) and older (population II) than the cepheids. They have periods from 10 to 30 days, and obey a similar period–luminosity relationship. They are sometimes called type II cepheids to distinguish them from the classical type I cepheids. *See* CEPHEIDS.

Pulsating stars. RR Lyrae are the second most common type of variable in the Galaxy. They have periods of less than 1 day, and have a small range of variation, from 0.5 to 1.5 magnitudes. They are particularly numerous in globular clusters, and thus are sometimes referred to as cluster variables. Their spectral classes range from A to F. The prototype of the class is RR Lyrae, which within 13.5 h varies from magnitude 7.0 to 8.1 and back. All RR Lyrae variables have the same intrinsic luminosity, of magnitude 0.5. Therefore these variables do not follow the period-luminosity relationship. Thus, if an RR Lyrae star can be identified in a star cluster and its apparent magnitude determined, the above equation and the star's known absolute magnitude can be used to obtain the distance to the cluster.

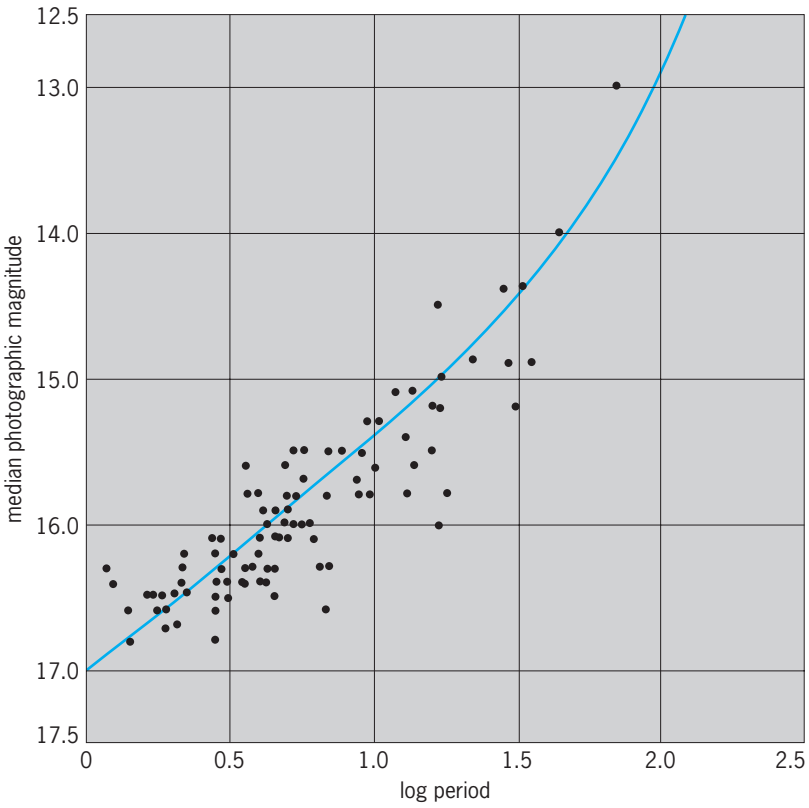


Fig. 2. Leavitt's period-luminosity curve for cepheid variables in Small Magellanic Cloud.

Long-period pulsating variable stars are the most abundant of all variables in the Milky Way Galaxy. They are red, cool, giant or supergiant stars with spectral class M or R, S, or C carbon types. Long-period variables are old stars which have evolved from the main sequence and are in the late stages of their evolution. Their size (diameter) is several hundred times that of the Sun; their mass, however, is about the same as the Sun. Their mass is concentrated in the core, which is surrounded by an extended gaseous atmosphere. *See* STELLAR EVOLUTION.

The light variability of long-period variables is due to the pulsation (contraction and expansion) of the star and its radiating surface. The prototype of long-period variables is o Ceti (Mira), which has given its name to those long-period variables that have a range of variation of 2.5 magnitudes and more. Mira variables have periods ranging from 100 to 1000 days. Although the change in brightness is periodic, the brightness of individual cycles may not be the same, some cycles being much brighter or fainter than others. Generally the rise to maximum brightness is faster than the decline to minimum brightness. *See* MIRA.

The spectra of long-period variables are abundant with lines of metallic oxides such as titanium oxide (TiO) and bright emission lines of hydrogen, seen particularly near maximum brightness. These bright lines are thought to be generated by propagating shock waves in the atmosphere of these stars.

Red giant and supergiant stars with less regularity of variation, shorter periods, and smaller ranges of variation, less than 2.5 magnitudes, are called semiregular variables. Other red variable stars that do not exhibit any regularity in their brightness change are called irregular stars.

Rare, very luminous, yellow supergiant stars that generally show alternating shallow and deep fadings are called RV Tauri stars. These pulsating variables vary by 2 to 3 magnitudes within 30 to 150 days.

Eruptive variables. Eruptive variables are those stars that have one or more eruptions—the ejection of matter into space—in their lifetime. These variables are sometimes referred to as cataclysmic variables because of the cataclysmic explosions they undergo. Cataclysmic variables are very close binary systems made up of an evolved, hot, dense white dwarf and a less evolved cool star which transfers mass onto the white dwarf accretion disk. *See* WHITE DWARF STAR.

There are many types of eruptive variables, ranging from supernovae and novae, in which spectacular explosions take place that brighten the system by many magnitudes, to nebular variables, which are young stars in the early stages of stellar formation, to R Coronae Borealis variables, which have sudden unpredictable decreases in their brightness. Only a few of the eruptive variable types will be discussed; a more complete listing of types is given in the table.

The most spectacular type of stellar explosion is a supernova, wherein a star's luminosity suddenly increases to tens of thousands of times the original brightness and the star outshines the total brightness of its galaxy. The gigantic explosion may be due to the gravitational collapse of a very hot and massive star that has exhausted the energy available from nuclear reactions; the collapse then creates enormous energy to blast outward the layers surrounding the stellar core. Alternatively, the explosion may occur in a close binary system in which one of the components is a massive white dwarf, an Earth-sized, very compact, old star, that explodes when it receives too much material from the other component. In the Milky Way Galaxy the most recent supernova observed was seen in 1604 in the constellation Serpens. However, Supernova 1987A, which exploded on February 23, 1987, in the Large Magellanic Cloud, a nearby companion galaxy to the Milky Way Galaxy, was visible to the unaided eye and was studied intensely. *See* SUPERNOVA.

The best known of the eruptive variables are stars that brighten by 7 to 16 magnitudes (less than 0.0010 of that of a supernova) within about a day. They stay at maximum brightness for a few days or weeks and then slowly fade. The word "nova," meaning new, was used for these stars. Actually a nova is not a new star, but an already existing star that due to the eruption has become very luminous and thus visible. *See* NOVA.

Recurrent novae are stars that about every 20 years or more have eruptions during which the system brightens by 7 to 9 magnitudes.

Dwarf novae have smaller-scale eruptions, in which the star brightens by 2 to 6 magnitudes within a day, stays bright 1 to 2 weeks, and then fades to the original brightness. The quasiperiodic eruptions may occur from ten to several hundred days apart. Dwarf novae are divided into two categories: U Geminorum and Z Camelopardalis stars. Z Camelopardalis stars are differentiated from U Geminorum variables by their periods of nonvariability midway between maximum and minimum brightness. These standstills can last days, months, or sometimes years.

R Coronae Borealis stars, instead of brightening through eruptions, irregularly decrease in brightness every 2 to 3 years by 1 to 9 magnitudes. These bright, highly luminous supergiant stars are rich in carbon and poor in hydrogen in composition. The decrease in brightness is caused by the veiling of the star by thick carbon clouds expelled to the star's atmosphere. *See* CATACLYSMIC VARIABLE; STAR.

Janet Akyüz Mattei; Arne Henden

Bibliography. G. North, *Observing Variable Stars, Novae and Supernovae*, 2004; C. Hellier, *Cataclysmic Variable Stars*, Springer, 2001; C. Hoffmeister, G. Richter, and W. Wenzel, *Variable Stars*, Springer, 1985; D. H. Levy, *Observing Variable Stars*, 2d ed., Cambridge University Press, 1998; J. R. Percy (ed.), *The Study of Variable Stars Using Small Telescopes*, Cambridge University Press, 1986; H. A. Smith, *RR Lyrae Stars*, Cambridge University Press, 1995; C. Sterken and C. Jaschek (eds.), *Light Curves of Variable Stars*, 2d ed., Cambridge University Press, 2005; B. Warner, *Cataclysmic Variable Stars*, Cambridge University Press, 1995.

Variational methods (physics)

Methods based on the principle that, among all possible configurations or histories of a physical system, the system realizes the one that minimizes some specified quantity. Variational methods are used in physics both for theory construction and for calculational purposes.

General principles. The earliest use of a variational principle for physics is Fermat's principle in optics, which states that when a light ray traverses a medium with nonuniform index of refraction its path is such as to minimize its travel time. The integral given in Eq. (1) expresses the time that the light takes to

$$T = \int_{\Gamma} \frac{ds\ n(s)}{c} \tag{1}$$

travel from point A to point B along a path Γ, where c is the speed of light in vacuum and n is the given index of refraction. An application of the calculus of variations to this integral makes it possible to determine the particular path Γ_0 for which T is a minimum. This problem is mathematically identical to the variational principle that determines a geodesic, the path of shortest distance, in a given geometry. In that form, the same principle determines the world lines of all objects in the general theory of relativity. *See* CALCULUS OF VARIATIONS; RELATIVITY; RIEMANNIAN GEOMETRY.

Similarly, in mechanics, Hamilton's principle for the action is defined for any system of point particles by Eq. (2). Here $L = K - V$ is the lagrangian

$$I = \int_{t_1}^{t_2} dt\ L \tag{2}$$

of the system, where K and V are its kinetic and potential energies, respectively, expressed in terms of the generalized coordinates and velocities of the particles. The integral extends over an arbitrarily prescribed path Γ in configuration space as a functions of the time t. Hamilton's principle asserts that the trajectories of all the particles are determined by the requirement that Γ be such that, for given initial and final times t_1 and t_2, the action I is a minimum; for this reason it is also called the principle of least action. If the calculus of variations is applied to implement this principle, the corresponding Euler-Lagrange equations are obtained—partial differential equations for the lagrangian, which express necessary and sufficient conditions for the trajectories to be such that the action is a minimum. These equations are the lagrangian equations of motion, in other words, Newton's equations of motion in lagrangian form. *See* ACTION; HAMILTON'S PRINCIPLE; LAGRANGE'S EQUATIONS; LAGRANGIAN FUNCTION; LEAST-ACTION PRINCIPLE; NEWTON'S LAWS OF MOTION.

The principle of least action has been generalized to infinitely many degrees of freedom, that is, fields. A Lagrange density function $\mathscr{L}$ is then defined, which is a function of the fields and their time derivatives at any given point in space and time. The action is given by Eq. (3), in which the x-integral extends over all

$$I = \int_{t_1}^{t_2} dt \int dx\ \mathscr{L} \tag{3}$$

three-dimensional space. The Euler-Lagrange equations of the principle of least action in this case are the field equations.

It is possible, in this way, to generate, for example, the Maxwell equations from the Lagrange density given in Eq. (4), where **E** and **H** are the electric

$$\mathscr{L} = \frac{1}{8\pi}(\mathbf{E}^2 - \mathbf{H}^2) + \frac{1}{c}\ \mathbf{j} \cdot \mathbf{A} - \rho\varphi \tag{4}$$

and magnetic fields respectively, **j** is the current density, **A** is the vector potential, ρ is the charge density, and φ is the scalar potential. *See* MAXWELL'S EQUATIONS; POTENTIALS.

Theory construction. For any field theory, only the Lagrange density needs to be given; the field equations are then derivable as the corresponding Euler-Lagrange equations. A similar technique makes it possible to derive the Schrödinger equation and the Dirac equation in quantum mechanics from specific Lagrange density functions. *See* NONRELATIVISTIC QUANTUM THEORY; QUANTUM MECHANICS; RELATIVISTIC QUANTUM THEORY.

This method has great procedural advantages. For example, it facilitates a check of whether the theory satisfies certain invariance principles (such as relativistic invariance or rotational invariance) by simply ascertaining whether the Lagrange density satisfies them. The corresponding conservation laws can also be derived directly from the lagrangian. In all modern quantum-field theories, it is customary to postulate the specific form of the theory by simply specifying the Lagrange density function; the variational principle then automatically leads to the field equations and implies all the conservation laws. *See* CONSERVATION LAWS (PHYSICS); QUANTUM FIELD THEORY; SYMMETRY LAWS (PHYSICS).

Calculational uses. The variational method plays an important role in quantum-mechanical calculations. For the computation of needed quantities in terms of functions that result from the solution of differential equations, it is always of great advantage to use formulas that have the special form required to make them stationary with respect to small variations of the input functions in the vicinity of the unknown, exact solutions.

For instance, the energy of a quantum-mechanical system may be expressed in terms of the expectation value of its hamiltonian, H, given by Eq. (5), with the

$$E = \frac{(\psi, H\psi)}{(\psi, \psi)} \tag{5}$$

wave function ψ set equal to the eigenfunction ψ_E of H at that energy. The expression in Eq. (5) has the property that it is stationary with respect to small variations of the function ψ in the neighborhood of ψ_E: Its first variation from $\psi = \psi_E$ vanishes. This implies that if, in the calculation, instead of the exactly correct eigenfunction ψ_E an approximation whose

norm differs from that of the correct eigenfunction by a small amount ϵ is used, then the calculated value of E will differ from the exact value by the much smaller amount ϵ^2. For the ground state of the system, that is, the state of lowest energy, this stationary value is, in fact, an absolute minimum. As a result, the number E calculated by Eq. (5) with any arbitrary function ψ is always an upper bound for the ground-state energy, and if ψ is a reasonable approximation, that bound will be close to the correct value. To find a similarly good approximation and bound for excited states, it is necessary to choose trial functions that are orthogonal to the ground state.

In practice, almost all computations of energy levels of atomic or nuclear systems are performed by using this variational method. A trial function depending on a number of parameters is adopted, and the parameters are chosen so that the calculated value of E is stationary; that is, its first derivatives with respect to the parameters all vanish. The value of E so obtained is the best approximation possible with the given set of trial functions. (This procedure is called the Ritz variational method.)

Analogous variational methods also exist for the calculation of other quantum-mechanical quantities, as well as for classical quantities that are eigenvalues of differential equations arising in dynamics. For example, the vibrational frequencies of mechanical systems can be calculated by such a method, as was done first by Lord Rayleigh. In quantum mechanics, there are several expressions for the scattering phase shift of a particle in a central potential that are stationary with respect to small variations of the scattering wave function near its exact form. R. Feynman used the variational method in the context of his path integrals to calculate the lowest energy of an electron in a polarizable crystal—a polaron. *See* FEYNMAN INTEGRAL; MINIMAL PRINCIPLES; POLARON.

Roger G. Newton

Bibliography. P. Blanchard, *Variational Methods in Mathematical Physics: A Unified Approach*, 1992; R. P. Feynman, Slow electrons in a polar crystal, *Phys. Rev.*, 97:660–665, 1955; P. Griffiths, *Exterior Differential Systems and the Calculus of Variations*, 1983; J. T. Oden and J. N. Reddy, *Variational Methods in Theoretical Mechanics*, 1983.

Varistor

Any two-terminal solid-state device in which the electric current I increases considerably faster than the voltage V. This nonlinear effect may occur over all, or only part, of the current-voltage characteristic. It is generally specified as $I \propto V^n$, where n is a number ranging from 3 to 35 depending on the type of varistor. The main use of varistors is to protect electrical and electronic equipment against high-voltage surges by shunting them to ground. *See* ELECTRIC PROTECTIVE DEVICES.

One type of varistor comprises a sintered compact of silicon carbide particles with electrical terminals at each end. It has symmetrical characteristics (the same for either polarity of voltage) with n ranging from 3 to 7. These devices are capable of application to very high power levels, for example, lightning arresters. *See* LIGHTNING AND SURGE PROTECTION.

Another symmetrical device, the metal-oxide varistor, is made of a ceramiclike material comprising zinc oxide grains and a complex amorphous intergranular material. It has a high resistance (about 10^9 ohms) at low voltage due to the high resistance of the intergranular phase, which becomes nonlinearly conducting in its control range (100–1000 V) with $n > 25$.

Semiconductor rectifiers, of either the *pn*-junction or Schottky barrier (hot carrier) types, are commonly utilized for varistors. A single rectifier has a nonsymmetrical characteristic which makes it useful as a low-voltage varistor when biased in the low-resistance (forward) polarity, and as a high-voltage varistor when biased in the high-resistance (reverse) polarity. Symmetrical rectifier varistors are made by utilizing two rectifiers connected with opposing polarity, in parallel (**Fig. 1***a*) for low-voltage operation and in series (Fig. 1*b*) for high-voltage use. For the high-voltage semiconductor varistor, n is approximately 35 in its control range, which can be designed to be anywhere from a few volts to several hundred. *See* SEMICONDUCTOR RECTIFIER.

Silicon rectifiers are fabricated mostly by solid-state diffusion techniques similar to those utilized for transistors and integrated circuits. A thin plate of *n*-type silicon is exposed to a gaseous atmosphere containing a dilute *p*-type impurity at a high temperature (about 2200°F or 1200°C). Some atoms of the *p*-type impurity diffuse into the silicon surface by displacing silicon atoms, forming a *pn* junction within about 0.001 in. (0.025 mm) of the surface.

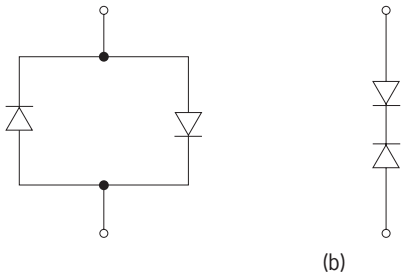


Fig. 1. Symmetrical rectifier varistors. (*a*) Low voltage. (*b*) High voltage.

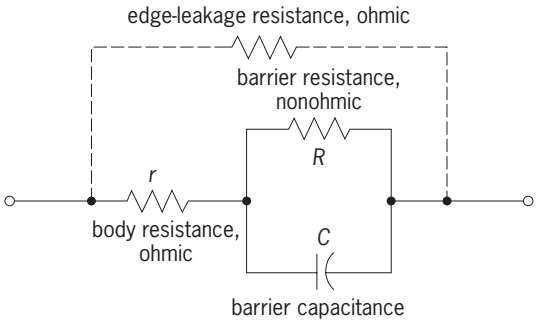


Fig. 2. Equivalent circuit of a single-rectifier varistor.

Metal electrodes are applied to *p*- and *n*-type regions by plating or evaporation. *See* INTEGRATED CIRCUITS; JUNCTION DIODE; TRANSISTOR.

The equivalent circuit of a *pn*-junction varistor is shown in **Fig. 2**. The capacitance is high, limiting high-frequency performance. I. A. Lesk

Bibliography. E. Bleuler and R. O. Haxby (eds.), *Methods of Experimental Physics*, vol. 2, pt. A, 1975; J. D. Harnden, Jr., et al., Metal-oxide varistor: A new way to suppress transients, *Electronics*, 45(21):91–95, October 9, 1972; L. M. Levinson (ed.), *Ceramic Transactions*, vol. 3: *Advances in Varistor Technology*, 1989; J. R. O'Connor and J. Smiltens (eds.), *Silicon Carbide*, 1960.

Varnish

A transparent surface coating which is applied as a liquid and then changes to a hard solid. Varnishes are solutions of resinous materials in a solvent, and dry by the evaporation of the solvent or by a chemical reaction, either with oxygen from the air or by some other means, including absorption of atmospheric moisture.

Spirit varnishes are those in which the evaporation of solvent is the only drying process; the solvent is usually alcohol, although the term is used for similar coatings made with other solvents. Shellac varnish, made by dissolving shellac in alcohol, is the most common of this type. Oleoresinous varnishes are made by treating a drying oil with a resin, usually with heat, and dissolving the reaction product in a solvent, usually a petroleum fraction; drying results from the evaporation of the solvent, followed by polymerization of the drying oil portion, a reaction which is accelerated by metallic driers added to the varnish. For a discussion of the mechanism of this drying action *see* DRIER (PAINT); SHELLAC.

Varnish coatings on wood are used to protect against abrasion, staining, and weather and to reduce the penetration of water and other materials without obscuring the grain or changing the color materially. Varnishes are used on masonry to reduce the penetration of moisture and the damage from freezing. Paper is coated with varnish to resist moisture and keep printing from being damaged. *See* PRINTING.

Shellac varnishes are used where fast-drying, slightly colored coatings of low cost are needed and where resistance to water, solvents, and weather is not required. Oleoresinous varnishes have greater resistance, but their color is usually darker. Spar varnishes are oleoresinous varnishes, often made from tung oil and a phenolic resin, which have good weather resistance; they were originally used for marine finishes.

Polyurethane varnishes exhibit superior abrasion resistance compared with other types. High gloss and gloss retention are other attractive features.

Varnish stains contain a dye dissolved in the solvent; they are used to change the color of a wood without obscuring the grain. *See* DYE; SOLVENT; SURFACE COATING. C. R. Martinson; C. N. Sisler

Varnish tree

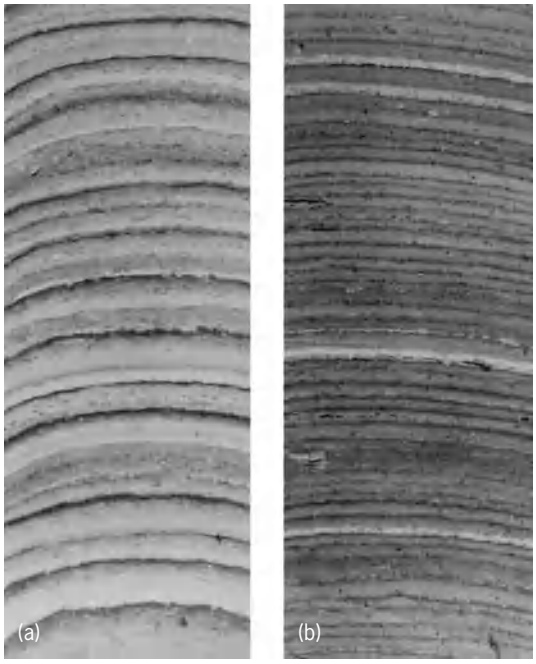
The plant *Toxicodendron vernicifluum* (previously known as *Rhus vernicifera*), also called lacquer tree, a member of the sumac family (Anacardiaceae). It is a native of China, but has long been cultivated in Japan. When the bark is cut, it exudes a milky juice which darkens and thickens on exposure. This is the lacquer long used in China and Japan. When properly applied, the thin transparent film becomes a varnish of extreme hardness. Lacquer is a remarkably protective coating as it is not altered by acids, alkalies, alcohol, or heat up to 160°F (71°C). Nut galls, iron in solution, and gold or other metals are mixed with the lacquer before drying to make the various kinds of lacquers. The process of lacquering is technical and tedious, sometimes requiring 300–400 coats and several years to complete the finish of one item. *See* LACQUER; SAPINDALES. Perry D. Strausbaugh; Earl L. Core

Varve

Any of a variety of distinct sediment laminations or beds deposited within the span of a single year. They are formed commonly in saline or fresh-water lakes, but examples from marine environments are known as well. Usually, varves occur in repetitive series and thus comprise vertical sequences of annual cyclic deposits. Varves range in thickness from less than a millimeter (0.04 in.) to over a meter (3 ft), but typically are a few millimeters or centimeters thick.

The classic varves are found in glacial lake sediments formed during the Pleistocene ice ages. These glacial varves occur typically as couplets of light-colored silt or sand and dark clay. The relatively coarse silt-sand layers are formed during the warm summer months when meltwater inflows and sediment yields to the lake are large. During winter when meltwater inflow is greatly reduced or stopped, the fine-grained clay settles slowly to the lake bottom to deposit the fine, dark winter layer of the varve. Similar varved sediments are found commonly in modern glacier-fed lakes that undergo large seasonal variations in inflow (see **illus.**). *See* GLACIATED TERRAIN; PLEISTOCENE.

Varves can be matched sometimes between sample localities and used as tools for correlation as well as for chronological reconstructions. Varve chronologies have been erected in late Quaternary deposits in numerous areas of northern Europe and North America formerly occupied by continental ice sheets. Such varve correlations and chronologies have been used to reconstruct the recession history of the last ice sheet as well as to establish an absolute time scale back to 10,000 years. In addition to their use in dating sedimentary deposits, varves have been used to investigate sedimentation rates, cyclic deposition, climate variations, glacial histories, and as standards of comparison for other dating techniques. Varves have been identified occasionally in ancient sedimentary rocks. *See* QUATERNARY.



Varved sediment in two lake-bottom cores from modern glacier-fed lake (Hector lake, Alberta, Canada). Varves collected closer to the river inlet (left) are thicker and represent higher annual sedimentation rates. The length of each sequence shown is 5 cm (2 in.).

Chemically precipitated varves, typically composed of calcium carbonate, have been recognized in numerous modern lakes and ancient lacustrine deposits. These arise from a variety of mechanisms, including seasonal fluctuations in temperature, salinity, and pH; variable water motions; and biological productivity cycles, especially seasonal diatom blooms which affect pH and, consequently, calcium carbonate solubility. Such varves commonly form in lakes that are chemically stratified (meromictic). *See* GLACIAL EPOCH; MARINE SEDIMENTS; MEROMICTIC LAKE; SEDIMENTARY ROCKS; SEDIMENTOLOGY; SHALE.

Norman D. Smith

Bibliography. R. Y. Anderson et al., Meromictic lakes and varved lake sediments in North America, *USGS Bull.*, 1607:1–19, 1985; I. Banerjee, Sedimentology of Pleistocene glacial varves in Ontario, Canada, *Geol. Surv. Can. Bull.*, 226A:1–44, 1974; E. M. Leonard, Varve studies at Hector Lake, Alberta, Canada, and the relationship between glacial activity and sedimentation, *Quaternary Res.*, 25:199–214, 1986; S. D. Ludlam, Fayetteville Green Lake, New York, III: The laminated sediments, *Limnol. Oceanog.*, 14:848–857, 1969; N. D. Smith, Sedimentation processes and patterns in a glacier-fed lake with low sediment input, *Can. J. Earth Sci.*, 15:741–756, 1978.

Vascular disorders

Those disorders that involve the arteries, veins, and lymphatics.

Arteries. The arteries distribute blood to the tissues. Those arteries with proper names, such as the aorta or the renal artery, are referred to as large and medium-sized arteries. They branch into vessels of smaller size, finally terminating in the smallest arteries, the arterioles. These microscopic vessels then connect with the capillaries, where the exchange of oxygen and carbon dioxide takes place. Diseases of the large and medium-sized arteries are the major cause of morbidity and mortality in the Western world.

Arteriosclerotic vascular disease. This disorder, also known as arteriosclerosis, affects large and medium arteries and is particularly common in the arteries supplying the heart, those supplying the brain, and those supplying the lower extremities. Progressive narrowing and finally total occlusion of these arteries lead to the development of angina pectoris, myocardial infarction, stroke, and vascular insufficiency in the limbs. For unknown reasons, the process rarely involves the arteries of the arm. In the legs, it leads to pain in the muscles with walking (intermittent claudication) and, in far advanced cases, gangrene. *See* GANGRENE; INFARCTION.

Although the cause of the disease is not known, certain factors appear to predispose to its development. These include cigarette smoking, hypertension, elevated serum cholesterol, and diabetes mellitus. There can also be a genetic predisposition to the disorder. There is increasing evidence that control of the above risk factors can prevent the onset of the disease or slow its progression. *See* DIABETES; HYPERTENSION.

The disorder appears to develop in stages, with early lesions sometimes seen even in young adults. The primary lesion is the plaque, which is an accumulation of fat, smooth muscle, and collagen that develops at branch points and bifurcations in the arterial system. As the plaque gets larger, it begins to impinge on the lumen of the artery, leading to a reduction in blood flow. In later stages, the plaque undergoes degeneration, with loss of the surface covering, the appearance of calcium, necrosis, and intraplaque hemorrhage. Portions of the plaque may break off, producing emboli that may block vital areas of the tissues supplied by the artery. In its final stages, clot formation (thrombosis) at the site of the plaque can occur, which completely shuts off the blood supply to an area of tissue. The extent to which the tissue survives then depends on the capability of accessory vessels that bypass the area of occlusion to take over the function of the occluded artery. *See* EMBOLISM; THROMBOSIS.

At some sites the arteriosclerosis may be accompanied by an aneurysm, a ballooning of the weakened vessel wall. Aneurysms most commonly occur in the abdominal aorta, where rupture is not uncommon. When they develop in the femoral and popliteal arteries of the lower limb, rupture rarely occurs but thrombosis is common. The forming thrombus within the aneurysm can break away and occlude the arteries of the lower limb and foot. *See* ANEURYSM.

Patients with diabetes mellitus are prone to develop arteriosclerosis in the coronary arteries, the

carotid artery in the neck, and the arteries supplying the lower limbs. For diabetics, the prevalence of arteriosclerosis is at least 20 times as high for the leg arteries and 8 times as high for the carotid artery compared with the general population. This high rate of occurrence in arteries supplying the limbs is the main reason for the very high rate of limb loss in persons with diabetes mellitus.

Although there is evidence that arteriosclerosis may be prevented or slowed by control of those risk factors associated with its occurrence, narrowing and occlusion of the large and medium arteries continues to be the major health problem in the Western world. To treat arteriosclerotic narrowings of vessels, surgery may be done to remove or bypass the obstruction, or a balloon may be placed into the artery and guided to the site of involvement, where it is inflated in order to compress the plaque. This procedure is known as balloon angioplasty. The goal is to restore normal blood flow to the involved tissues. *See* ARTERIOSCLEROSIS.

Inflammation. Much less common are diseases that lead to arteritis, an inflammation of the large or medium arteries, with subsequent occlusion or rupture of the artery. The causes include bacterial infection, syphilis, allergic disorders, and hypersensitivity states.

Lodging of bacteria in the wall of an artery can lead to development of a mycotic aneurysm. The bacteria originate from a remote site in the body where they gain access to the circulation. It is not known why they localize within the artery wall, but when such localization occurs, the artery can become weakened, leading to development of an aneurysm that may rupture or be the site of further dissemination of the bacteria throughout the body. Mycotic aneurysms can develop, for example, in patients with an infection involving one of the valves of the heart (bacterial endocarditis).

Before penicillin was introduced, syphilis was a common cause of infection of the arterial wall. It led to destruction of the elastin and smooth muscle of large arteries such as the aorta. The wall of the aorta was ultimately weakened to the point where an aneurysm would occur. This disease is rare in the United States today presumably because of the effectiveness of penicillin in killing the causative organism. *See* SYPHILIS.

The hypersensitivity disorders that can lead to arteritis include periarteritis nodosa, lupus erythematosus, rheumatic arteritis, temporal arteritis, and giant-cell arteritis. When these affect the arterial wall, thrombosis of the involved artery is the most common outcome. The thrombosis occurs gradually and leads to symptoms reflecting the chronic reduction in blood flow to the organ. The clinical manifestations depend on the artery involved and the tissues supplied. *See* CONNECTIVE TISSUE DISEASE; HYPERSENSITIVITY.

Small arteries. Several generalized disorders can involve the small arteries. These include scleroderma, periarteritis nodosa, lupus erythematosus, rheumatoid arthritis, and dermatomyositis. The involvement can produce areas of cell death (necrosis) and ulceration of the skin, particularly of the limbs. This appears to occur because of the occlusion of the involved small arteries, with consequent loss of blood supply. Such involvement often is preceded by a type of hypersensitivity to cold, a condition known as Raynaud's syndrome. Such hypersensitivity can also occur when the small arteries and arterioles go into spasm when the hands or feet are exposed to cold. When this occurs, the digits, particularly the fingers, turn white, then blue, and finally red. This has been referred to as the triphasic color response. When it is not associated with an underlying disease such as scleroderma, it poses no risk. However, if it is associated with an underlying disorder, the small arteries of the hands and feet may become occluded, and digit ulcers or gangrene can develop.

Veins. The most common disorder of the venous circulation is varicose veins. It develops because of the loss of function of the valves in the superficial veins of the limbs. The two most commonly involved veins are the greater and lesser saphenous. In this condition, the veins are enlarged and often tortuous. If the deep veins are not involved, the varicose veins are termed primary. Varicose veins tend to cluster in families and are more common in women than men. The condition may be worsened by pregnancy due to both hormonal effects and an increase in the venous pressure in the iliac veins due to the enlarging uterus. Primary varicose veins may be cosmetically unpleasing, but they rarely cause serious symptoms in the involved limbs.

A more serious problem develops when the varicose veins are associated with destruction of valves in the deep veins, secondary to deep-vein thrombosis. Blood then is forced to flow in a retrograde direction and through the perforating veins to join the superficial veins. As a result, the superficial veins become enlarged and tortuous. Because of the combined involvement of the deep and superficial veins, leg swelling and pain are common. As time passes, brownish discoloration of the skin of the lower leg can occur. In addition, ulceration is common and can be difficult to treat. When varicose veins develop because of problems within the deep venous system, they are called secondary varicose veins.

When patients undergo major surgery or are confined to bed because of a serious medical illness, thrombosis of the veins of the leg can occur. This is commonly referred to as phlebitis but this term is not correct since inflammation is not a prominent part of the process. It is more appropriate to refer to this disorder as deep-vein thrombosis. The thrombosis occurs most commonly in the deep veins of the calf but can progress and extend proximally to involve the popliteal, femoral, and iliac veins. When this occurs, portions of the thrombi can dislodge, leading to the development of emboli in the lung. Emboli that arise from veins as large as the femoral or iliac may be quite large and can prove fatal.

Lymphatics. The lymphatics are responsible primarily for draining interstitial fluid from the tissues

in which they are found. This fluid, which is rich in protein, must be removed, or swelling (edema) will occur. When this drainage function is disturbed by obstruction of the lymphatics by tumors, parasites, or surgical excision, serious swelling can occur. This condition is called lymphedema. The pattern of swelling differs from that seen with obstruction of the venous system. For example, it will involve not only the lower leg but the foot and toes as well. In lymphedema, it is not uncommon for infection with a betahemolytic streptococcus to occur. When this happens, the patient develops shaking chills, fever, and pain in the limb. Prompt treatment with an antibiotic such as penicillin is necessary to avoid further destruction of the lymphatics and worsening of the edema. *See* EDEMA; LYMPHATIC SYSTEM.

Obstruction of the lymphatics can also be congenital or can appear unexpectedly at the time of the onset of the menses. The mechanisms are unknown, but the effects are the same as when a definite cause is found. Regardless of its cause, lymphedema is a permanent condition. Its control requires the use of compression stockings, which limit the amount of fluid that can accumulate in the legs during the day when the individual is upright.

Arteriovenous communications. Proper function of the cardiovascular system occurs when blood follows its normal course via the arteries to the capillaries and veins and then back to the right heart. If there are anomalous communications between the arteries and veins proximal to the capillaries, blood is shunted away from the capillary bed. Such shunting can be congenital or acquired.

The congenital form can occur at almost any site in the vascular system. When these communications are present, they may appear as prominent veins that are much larger and more numerous than normal. These are commonly referred to as hemangiomas. The extent to which they cause problems depends upon their location, their size, and numbers of feeding communications that are present. When they occur in the brain, the thin-walled and enlarged veins can rupture leading to intracranial hemorrhage, which may be fatal as the intracranial pressure rises often to very high levels. When they occur in the limbs, lungs, or abdomen, the problems are most often related to the amount of blood that is shunted and thus bypasses the capillary network. If large volumes of blood are shunted, there may be an extra load on the heart, leading to cardiac hypertrophy and ultimately to heart failure and death. Correction of these conditions is difficult or impossible if the malformations are extensive.

If a foreign body such as a bullet or knife passes between a major artery and vein, an artificial communication can develop that results in the shunting of a large volume of blood. When the communication is between very large vessels such as the abdominal aorta and inferior vena cava, the amount of blood shunted may be so large that heart failure occurs. These abnormal communications require surgical correction to restore normal blood flow. *See* CARDIOVASCULAR SYSTEM. D. Eugene Strandness, Jr.

Bibliography. W. R. Hazzard, R. Andres, and E. L. Bierman (eds.), *Principles of Geriatric Medicine*, 4th ed., 1998; E. Braunwald et al. (eds.), *Harrison's Principles of Internal Medicine*, 15th ed., 2001.

Vector (mathematics)

A mathematical element used to describe quantities having magnitude and direction. It can be represented by a directed line segment. Vectors are widely used in mathematical physics. *See* CALCULUS OF VECTORS. Marvin Yelles

Vector methods (physics)

Methods that make use of the behavior of physical quantities under coordinate transformations.

Definitions. From the point of view of physics, the most appropriate definition of a vector in three-dimensional space is a quantity that has three components which transform under rotations of the coordinate system like the coordinates of a point in space. What characterizes rotations is that the distance from the origin, $\sqrt{x_1^2 + x_2^2 + x_3^2}$ of all points $\mathbf{x}$ with cartesian coordinates x_i, $i = 1, 2, 3$, remains unchanged. Specifically, if the rotation takes the x_i to new coordinates x_i', given by Eqs. (1) [in which

$$x_i' = \sum_{j=1}^{3} a_{ij} x_j \qquad i = 1, 2, 3$$
$$\sum_{k=1}^{3} a_{ik} a_{jk} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \qquad \det\{a_{ij}\} = 1 \tag{1}$$

$\det\{a_{ij}\}$ is the determinant of the matrix $\{a_{ij}\}$], then the three quantities V_i, $i = 1, 2, 3$, form the components of a vector $\mathbf{V}$, if in the new coordinate system the transformed coordinates are given by Eq. (2).

$$V_i' = \sum_{j=1}^{3} a_{ij} V_j \qquad i = 1, 2, 3 \tag{2}$$

If the coordinate transformation of Eqs. (1) is such that $|a_{ij}| = 1$ for $i = j$ and $a_{ij} = 0$ for $i \neq j$, but such that $\det\{a_{ij}\} = -1$ rather than $+1$ as in Eqs. (1), then it describes a reflection, in which a right-handed coordinate system is replaced by a left-handed one. [Transformations that combine Eqs. (1) with a reflection are called improper rotations, whereas Eqs. (1) is called a proper rotation.] If, for such a transformation, Eq. (2) also holds, then $\mathbf{V}$ is called a polar vector, whereas if the components of $\mathbf{V}$ do not change sign, it is called an axial vector or pseudovector. The vector $\mathbf{V}$ can also be looked upon as a quantity with a direction, with the magnitude $\sqrt{V_1^2 + V_2^2 + V_3^2}$ pointing from the origin of the coordinate system to the point in space with the cartesian coordinates (V_1, V_2, V_3).

A quantity that remains invariant under a rotation of the coordinate system is called a scalar. The dot product of two vectors, given by Eq. (3), is a scalar,

and the vector product, given by Eq. (4), where ϵ_{ijk}

$$\mathbf{V} \cdot \mathbf{W} = \sum_{i=1}^{3} V_i W_i \tag{3}$$

$$(\mathbf{V} \times \mathbf{W})_i = \sum_{j,k=1}^{3} \epsilon_{ijk} V_j W_k \tag{4}$$

is antisymmetric under exchange of any two of its indices, and $\epsilon_{123} = 1$, is an axial vector if both $\mathbf{V}$ and $\mathbf{W}$ are polar. *See* CALCULUS OF VECTORS.

The importance of vectors and scalars in physics derives from the assumed isotropy of the universe, which implies that all general physical laws should have the same form in any two coordinate systems that differ only by a rotation. It is therefore useful to classify physical quantities according to their transformation properties under coordinate rotations. Examples of scalars include the mass of an object, its electric charge, its volume, its surface area, the energy of a system, and its temperature. Other quantities have a direction and thus are vectors, such as the force exerted on a body, its velocity, its acceleration, its angular momentum, and the electric and magnetic fields. Since the sum of two scalars is a scalar, and the sum of two vectors is a vector, it is important in the formulation of physical laws not to mix quantities that have different transformation properties under coordinate rotations. The sum of a vector and a scalar has no simple transformation properties; a law that equated a vector to a scalar would have different forms in different coordinate systems and would thus not be acceptable.

Examples. Newton's second law, Eq. (5), in which

$$\mathbf{F} = m\mathbf{a} \tag{5}$$

$\mathbf{F}$ is the force on a particle, m is its mass, and $\mathbf{a}$ is its acceleration, has the same form in any two coordinate systems that are rotated with respect to one another. This is because both $\mathbf{F}$ and $\mathbf{a}$ are vectors, and m is a scalar; therefore $\mathbf{F} - m\mathbf{a}$ is a vector, and if it vanishes in one coordinate system (that is, all three of its components vanish), then it also vanishes in any rotated system. *See* NEWTON'S LAWS OF MOTION.

If the dot product of Newton's second law is taken with the velocity $\mathbf{v}$, the result, givenby Eq. (6),

$$\mathbf{v} \cdot \mathbf{F} = m\mathbf{v} \cdot \mathbf{a} \tag{6}$$

equates scalar quantities. The left-hand side of this equation is the rate at which work is being done, and the right-hand side is the time-derivative of $\frac{1}{2}m\mathbf{v}^2$, which is the kinetic energy. Therefore, this equation expresses the conservation of energy in this case: the rate at which work is done on the particle equals the rate at which its kinetic energy increases. *See* CONSERVATION OF ENERGY; DIFFERENTIATION; ENERGY.

Similarly, if the vector product of Newton's second law with the position vector $\mathbf{x}$ of the particle is taken, the result equates two polar vectors. The left-hand side is the torque applied (about the origin of the coordinate system), and the right-hand side is the rate at which the angular momentum (with respect to the origin) increases. Thus the conservation of angular momentum for the particle is obtained. *See* ANGULAR MOMENTUM; RIGID-BODY DYNAMICS; TORQUE.

Maxwell's equations for the electric and magnetic fields, $\mathbf{E}$ and $\mathbf{B}$, respectively, in vacuum, are Eqs. (7),

$$\begin{aligned} \nabla \times \mathbf{B} &= \frac{4\pi}{c}\mathbf{j} + \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} & \nabla \times \mathbf{E} &= -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} \\ \nabla \cdot \mathbf{E} &= 4\pi\rho & \nabla \cdot \mathbf{B} &= 0 \end{aligned} \tag{7}$$

in which $\mathbf{j}$ is the current density, ρ is the charge density, and c is the speed of light. These are vector equations, and hence are form-invariant with respect to coordinate rotations. Since the curl is a vector and the divergence is a scalar, the first two are vector equations and the second two are scalar. However, since ρ is a scalar, the third equation implies that $\mathbf{E}$ is a polar vector, and the second implies that $\mathbf{B}$ has to be an axial vector: its components do not change sign under a coordinate reflection. *See* MAXWELL'S EQUATIONS.

Generalizations. There are other physical quantities that transform like tensors under coordinate rotations. Indeed, axial vectors are really antisymmetric tensors of rank 2. Since an antisymmetric 3×3 matrix has exactly three independent entries, such a tensor in three dimensions has three components that transform under proper rotations like those of a vector. The stress tensor is an example of a physical quantity that transforms like a symmetric tensor of rank 2. *See* TENSOR ANALYSIS.

When measurements of an event in one laboratory are compared with those of the same event as performed in another laboratory moving uniformly with respect to the first, the Lorentz transformation of the special theory of relativity transforms simultaneously four quantities, the three spatial coordinates and the time. It is therefore useful to combine these four into a four-dimensional space called Minkowski space. The requirement of the special theory of relativity, that all physical laws should have the same form in any two laboratories in uniform motion with respect to one another, can consequently be most simply implemented by classifying physical quantities according to their transformation properties under Lorentz transformations, and demanding that laws should never equate quantities which transform differently. If that demand is met, the law will automatically be form-invariant with respect to such transformations. *See* LORENTZ TRANSFORMATIONS; RELATIVITY.

For that reason, the concept of a vector has been generalized to four dimensions: a four-vector is a set of four quantities that transform under Lorentz transformations just like the spatial coordinates (x_1, x_2, and x_3) and the time (t). There is, however, one important difference between the Minkowski space in which the relativistic four-vectors live and the three-dimensional coordinate space: the square of the distance that remains invariant under Lorentz transformations is $x_1^2 + x_2^2 + x_3^2 - c^2t^2$, which is required so that the speed of light c be the same for all observers. That makes the Minkowski space

noneuclidean. Similarly, the concept of a tensor is generalized to Minkowski space. *See* NONEUCLIDEAN GEOMETRY.

An example of a physical four-vector is the four-momentum of a particle, the first three components of which are its ordinary momentum (**p**), and the fourth, its energy (E) divided by c. Its invariant length is given by Eq. (8), which expresses the relativistic

$$\sqrt{E^2/c^2} - p \cdot p = mc \tag{8}$$

relation between the momentum of a particle and its energy. The form invariance of the Maxwell equations is most easily exhibited by combining the electric and magnetic fields into an antisymmetric field tensor of rank 2. The Maxwell equations then simply state that the divergence of this tensor is equal to the four-dimensional current density (which combines the three-dimensional current and the charge density), and that the divergence of its dual vanishes. *See* RELATIVISTIC ELECTRODYNAMICS.

In their most modern guise, vector methods make use of the coordinate-free concepts of exterior forms and the exterior differential calculus.

Roger G. Newton

Bibliography. R. L. Bryant, *Exterior Differential Systems*, 1991; D. A. Danielson, *Vectors and Tensors in Engineering and Physics*, 2d ed., 1997; F. A. Hinchey, *Vectors and Tensors for Engineers and Scientists*, 1976.

Vega

The fifth brightest of all stars and the third brightest in the northern sky. It will be the north polar star in about 12,000 years. In moving through the Milky Way Galaxy, the Sun is generally heading toward the position now occupied by Vega. At a distance of 7.8 parsecs (25.3 light-years, or 2.4×10^{14} km, or 1.49×10^{14} mi), Vega, or α Lyrae, is the prototypical star of spectral class A0V, indicating that it has an effective surface temperature of 9600 K (16,800°F) and derives its energy from the thermonuclear burning of hydrogen in a stable core region. Stars of this class have main-sequence lifetimes of about 5×10^8 years, 20 times shorter than the Sun. Vega's spectrum shows strong features due to hydrogen in the star's outer layers. The sharpness of these features implies that Vega is pointing its rotation axis nearly at the Sun. In comparison with the Sun, Vega is approximately 2.9 times larger in diameter, 2.5 times more massive, and 60 times more luminous. *See* PRECESSION OF EQUINOXES; SPECTRAL TYPE.

In most respects Vega is a typical A-type star made noteworthy by its proximity to the Sun. Two additional aspects of Vega are notable. First, unlike most stars of its type, Vega has no stellar companions gravitationally bound to it to form a binary star system. Second, Vega emits far more radiation at infrared wavelengths than would be expected. This radiation originates from a shell or disk of particles with a temperature of 100 K (−280°F) surrounding Vega out to a distance of 1.3×10^{10} km (8×10^9 mi), twice the radius of the solar system. The shell particles are at least 1 mm (0.4 in.) in size and are likely to be left over from Vega's formation. Near its edge, this dusty disk exhibits two prominent clumps of material opposite each other that may be the result of a gravitational resonance with a planet orbiting around Vega about midway through the disk. No such planet has yet been discovered, however. *See* INFRARED ASTRONOMY; STAR.

Harold A. McAlister

Bibliography. A. Fraknoi, D. Morrison, and S. Wolff, *Voyages Through the Universe*, 3d ed., Brooks Cole, 2004; I. Heinrichsen, H. J. Walker, and U. Klaas, Infrared mapping of the dust disk around Vega, *Monthly Notices Roy. Astron. Soc.*, 293:L78–L82, 1998; J. B. Kaler, *The Hundred Greatest Stars*, Springer, 2002.

Vegetable ivory

The seed of the tagua palm (*Phytelephas macrocarpa*) of tropical America. Each drupelike fruit contains six to nine bony seeds. The extremely hard endosperm of the seed is used as a substitute for ivory. Vegetable ivory can be carved and tooled to make buttons, chess pieces, knobs, inlays, and various ornamental articles. The hard, white seeds (coquilla nuts) of the Brazilian palm (*Attalea funifera*) are often used as a substitute for vegetable ivory in making the same or similar articles. *See* ARECALES.

Perry D. Strausbaugh; Earl L. Core

Vegetation and ecosystem mapping

The graphic portrayal of spatial distributions of vegetation, ecosystems, or their characteristics. Vegetation is one of the most conspicuous and characteristic features of the landscape and has long been a convenient way to distinguish different regions. Because vegetation provides the basic framework of terrestrial ecosystems, maps of ecosystems and biomes have been mainly vegetation maps. Resources generally must be inventoried and mapped before they can be well managed. Thus, as pressure on the Earth's natural resources grows and as natural ecosystems are increasingly disturbed, degraded, and in some cases replaced completely, the mapping of vegetation and ecosystems, at all scales and by various methods, has become more common and important. *See* BIOME; ECOSYSTEM.

Basic approaches and considerations. Probably since classical times, and certainly since the rapid development of mapmaking during the Renaissance, mappers have located vegetation in relation to natural or cultural patterns. Three approaches arose for mapping general vegetation patterns, (1) based on vegetation structure or gross physiognomy, (2) based on correlated environmental patterns, and (3) based on important floristic taxa. The environmental approach provides the least

information about the actual vegetation but succeeds in covering regions where the vegetation is poorly understood. While most earlier and many later vegetation maps involve combinations of the three approaches, most modern classification systems use a combination of physiognomic and floristic characters. *See* PLANT GEOGRAPHY.

The first world vegetation maps were produced in Europe near the end of the nineteenth century and were inconsistent since many regions of the world were not well known. As knowledge of obscure regions increased, better maps appeared for the whole world and for individual continents, countries, and regions. Perhaps the first world vegetation-based map free of environmental surrogates was that by A. W. Küchler in 1949. Excellent regional vegetation maps include those of the United States by Küchler (1964), South America by Hueck and Seibert (1972), the Mediterranean and adjacent regions by UNESCO (1969), Western Australia by Beard, and France and India by the French vegetation mapping groups at Toulouse, Grenoble, and Pondicherry (India). In the United States, excellent vegetation maps have also appeared for many states, with particularly detailed versions for California, Minnesota, and Wisconsin.

Mapping has expanded since the 1950s to involve other aspects of vegetation and ecosystems as well as new methodologies for map production. Functional processes such as primary production, decomposition rates, and climatic correlates (such as evapotranspiration) have been estimated for enough sites so that world maps can be generated. Structural aspects of ecosystems, such as total standing biomass or potential litter accumulations, are also being estimated and mapped. Quantitative maps of these processes or accumulations can be analyzed geographically to provide first estimates of important aspects of world biogeochemical budgets and resource potentials.

Computer-produced maps, using Geographic Information Systems (GIS), often coupled directly with predictive models, remote-sensing capabilities, and other techniques, have also revolutionized vegetation and ecosystem mapping. This gives scientists a powerful tool for modeling and predicting the outcome from global climate change, in that feedback from the world's vegetation can be accounted for. Before computer technology exploded in the early 1980s, the spatial scale and related resolution or grain of vegetation and ecosystem mapping was limited by the static nature of hard-copy maps. The advent of GIS technology enabled the analysis of digital maps at any spatial scale, and the only limitation was the resolution at which the data were originally mapped. In addition, GIS software is used for sophisticated spatial analyses on maps, and this was virtually impossible before. *See* CLIMATE MODELING; GEOGRAPHIC INFORMATION SYSTEMS.

Spatial scale and base maps. The concept of spatial scale is relevant to many fields, including geography, cartography, and ecology. Spatial scale is often confused with the term "resolution" or "grain" because static maps have a set scale and resolution (the smallest object that can be seen on the map). With a GIS-based map, one can zoom in and out electronically, adjusting the scale and the resolution simultaneously. The only limitation is the resolution or grain of the original electronic map. Consider, for example, an electronic map of Washington State in which the original resolution of the map is very high, that is, features as small as 10 m (33 ft) across were measured. If the whole state is in view (that is, small-scale representation), the small-sized features are not visible. However, if one zooms in on a city (that is, large-scale representation), the outlines of individual buildings will be seen.

In terms of nomenclature, a 1:24,000 scale map is large-scale, and a 1:1,000,000 scale map is small-scale. This is often reversed, which is incorrect.

The U.S. Geological Survey large-scale maps (at 1:24,000) may be too coarse for working with plant associations and diverse cultural features. They are widely available, however, and can be enlarged either photostatically or using GIS software for field work. U.S. Army Map Service sheets at 1:250,000 are available in a consistent style over wide areas. If maps of these two types have not become too outdated because of land clearing, green forest overlay printing on these maps will provide a first-order distinction between forest/woodland and other kinds of vegetation. Such maps at least can serve to show altitudinal zonation, and may further suggest relations to geologic substrate and soil-type associations if reliable soil maps can be found. *See* CARTOGRAPHY; GEOLOGY; GEOMORPHOLOGY; MAP DESIGN; MAP REPRODUCTION.

The mapping of large areas at coarser scales, even with the benefit of the remote-sensing techniques mentioned below, is of varying accuracy for regions with different economic and landscape conditions. A common denominator of mapping at a widely available and acceptable scale convention of 1:1,000,000 has been advocated.

The following hierarchy, the U.S. National Vegetation Classification (USNVC) system, was developed jointly by over a dozen federal agencies, The Nature Conservancy (TNC), and the Ecological Society of America. It is used for classifying terrestrial ecological communities, based primarily on their vegetation.

System (for example, Terrestrial)

Physiognomic levels
- Formation Class
  - Formation Subclass
    - Formation Group
      - Formation Subgroup
        - Formation

Floristic levels
- Alliance
- Association

The classification system is internationally recognized, and it builds upon the UNESCO (1973) classification by incorporating the classification scheme

Summary of the classification hierarchy by the Association for Biodiversity Information

Level	Primary basis for classification	Level divisions and examples
Class	Type, height, and relative percentage of cover of the dominant, uppermost vegetation	Seven classes: Forest, Woodland, Shrubland, Dwarf-shrubland, Herbaceous, Nonvascular, and Sparse Vegetation
Subclass	For Forest, Woodland, Shrubland, and Dwarf Shrubland classes: leaf character	Three subclasses in each: evergreen, deciduous, and mixed evergreen-deciduous (no mixed evergreen-deciduous, dwarf-shrubland subclasses have been defined)
	For Herbaceous class: persistence and growth form	Four subclasses: perennial grasslands, perennial forb vegetation, annual grass and forb vegetation, and hydromorphic vegetation
	For Nonvascular class: relative dominance of nonvascular vegetation type	Three subclasses: lichens, mosses, algae
	For Sparse Vegetation class: particle sizes of the substrate features	Three subclasses: consolidated rock; boulder, gravel, cobble, or talus; and unconsolidated material (soil, sand, or ash)
Group	Varies by class: leaf characteristics, broad climatic types, presence and character of woody strata, major topographic position types or landforms	About 60 groups; example: Temperate or Subpolar Needle-leaved Evergreen Forest
Subgroup	Relative human impact	Two subgroups: Natural/Semi-natural or Cultural
Formation	Additional structural and environmental factors, including hydrology	Many; example: Saturated Temperate or Subpolar Needle-leaved Evergreen Forest
Alliance	Dominant/diagnostic species, usually of the uppermost or dominant stratum	Many; example: *Picea mariana* Saturated Forest Alliance
Association	Additional dominant/diagnostic species from any strata	Many; example: *Picea mariana/Alnus incana/Sphagnum* Forest

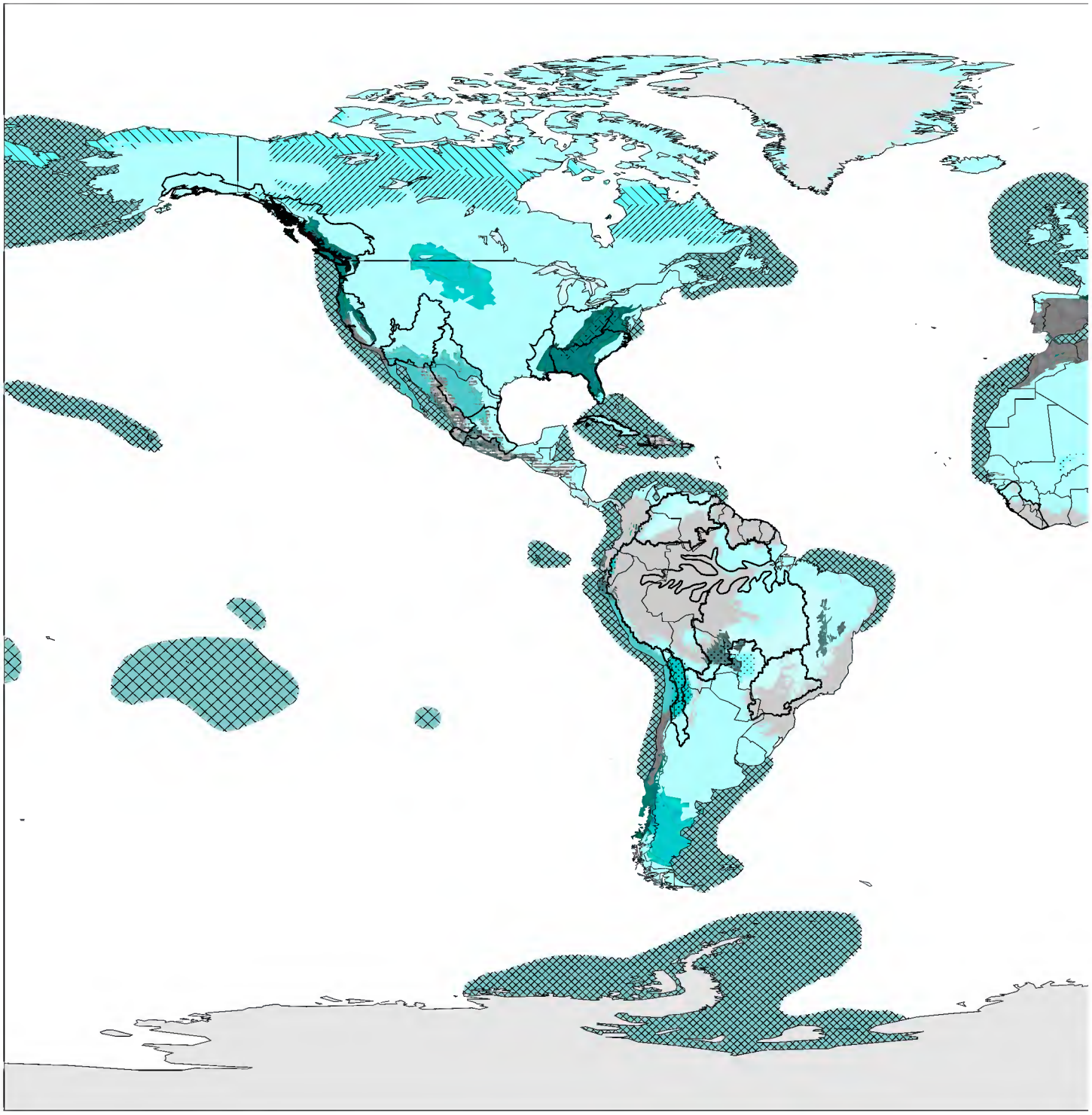
of R. S. Driscoll et al. (1984), and by increasing detail with the addition of alliance and association floristic levels. It is essentially identical to the National Vegetation Classification Standard (NVCS, FGDC 1997). This system also includes global rank definitions of conservation status ranking, making it an extremely powerful tool for identifying and prioritizing communities for conservation. However, worldwide maps have not yet been produced from this classification system. The NVCS classification hierarchy is summarized in the **table.**

For large countries, wall maps with good color printing can show many details by using patterns and symbols at smaller scales. Again, GIS-based maps are not necessarily limited this way. For the United States, Küchler's 1:3,168,000 map of 116 kinds of "Potential Natural Vegetation" has considerable detail related to patterns of rivers, which are shown on the map, and mountains, which are not shown but are readily recognized by the distribution of montane conifer or meadow types. For the former Soviet Union, an eight-sheet wall map at 1:4,000,000 shows even finer detail for the actual vegetation, with 109 colored mapping units, many of them subdivided by letters and supplemented by symbols for significant species. Coarser scales were required (varying with continent and type of map) in the Soviet Union's "Physical-Geographic Atlas of the World," but there was remarkably little loss of detail.

Colors are generally required for good maps of vegetation types, but process and structural quantities can usually be shown best by black-and-white shading. Computer-printed GIS maps are easily edited, so a series of maps can be rapidly produced that represent enormous amounts of information. In addition, sophisticated spatial analyses can be done using GIS software that are not possible using conventional static maps.

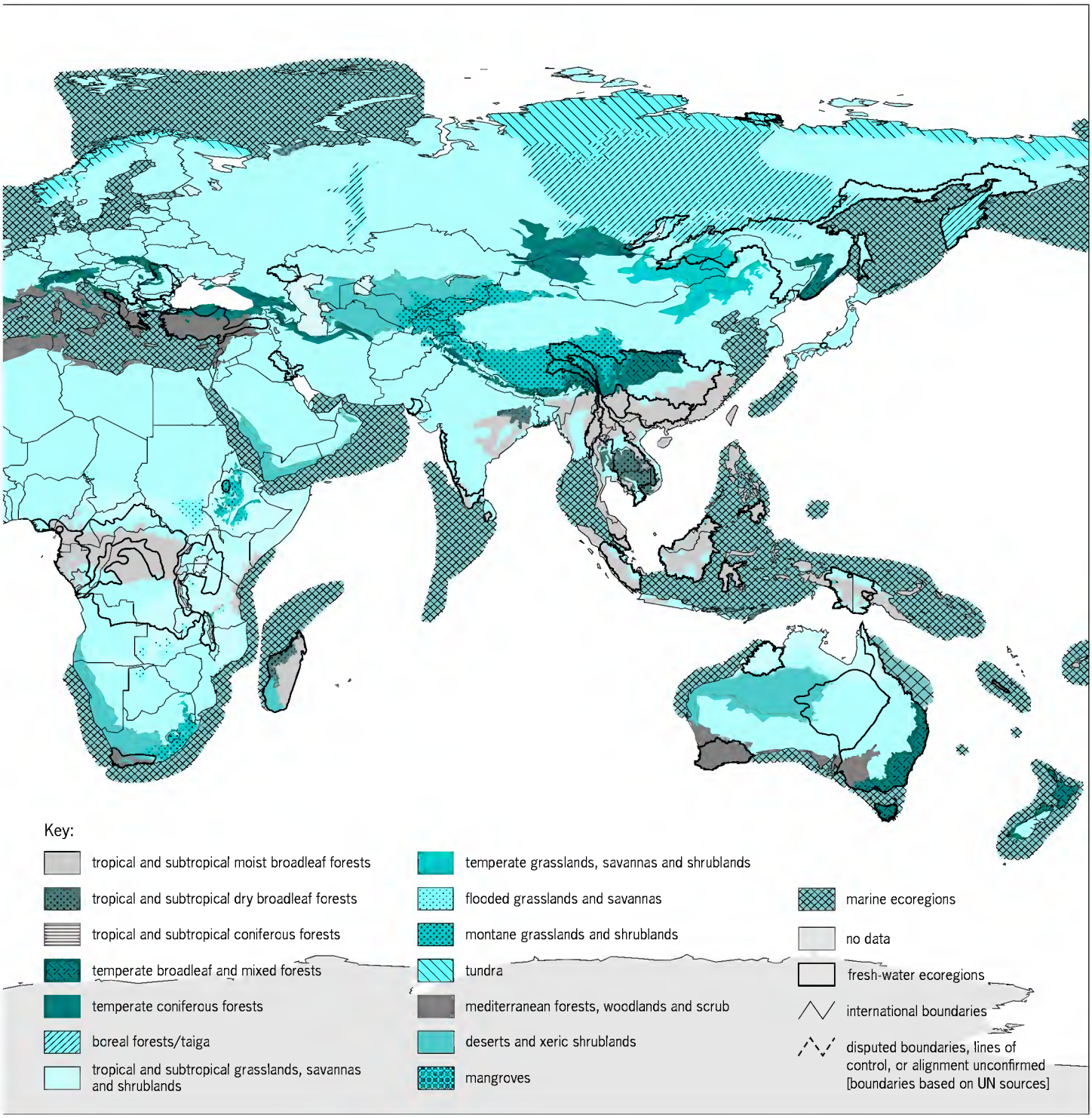
Topographic base maps may be improved to show such features as slope and aspect (direction), as well as altitude, in more detail. Simple, pale contour lines provide this information without distracting from the markings superimposed to represent vegetation or ecosystems. Topographic shading makes relief and drainage stand out more sharply but may interfere with the biological patterns.

Data gathering. Two basic approaches to detailed local vegetation analysis are (1) classification, which is usually based on intensively sampled plots, called relevés or quadrats, and (2) ordination, which usually involves some form of gradient or continuum analysis. Ordination is a method for plotting and



Terrestrial major habitat types.

inferentially ordering sampled vegetation plots in a usually hypothetical environmental space based on the relative similarity or dissimilarity of plots to each other. Both approaches require detailed field sampling and may yield data and patterns which are not easily mapped. More generalized mapping that focuses on vegetation types and other ecological patterns over larger areas requires more extensive data gathering, which can be obtained from remote-sensed imagery. Remote-sensed imagery is any imagery that is collected from a remote location, such as an aircraft or a satellite. There are two types of remote-sensed imagery: aerial photography and digital imagery. Aerial photography uses the familiar glass lens and photosensitive film technology. It is usually deployed from aircraft, but satellite deployment is also used. Digital imagery may or may not require a lens, but the photosensitive film has



been replaced by electronic sensors that send their information to computers, where the data are stored and analyzed. Digital imagery has typically been deployed from satellites, but with improvements in technology, aircraft-deployed sensors are becoming more common. *See* AERIAL PHOTOGRAPH.

Aerial photography. Aerial photography has been used since the 1920s to provide direct images of forests and other visibly distinct vegetation cover, with or without extensive ground checking to provide species recognition and other calibration data. This has been especially useful in countries with large, sparsely populated areas. Large areas are stereoscopically covered from high altitudes for surveys, road and harvest plans, and fire control operations. In addition, very fine photographs have been made of statistically selected plots, permitting measurement of shadow and tree height and crown form to provide

a basis for estimating tree volume and biomass. *See* PHOTOGRAMMETRY.

Soil surveys are also made from aerial photographs, such as to estimate soil moisture or to measure areas for which conservation payments are made to individual farmers. But even when land is freshly plowed, there are limits to what can be inferred from the air or from observation of surface soil horizons. Local experience and "ground-truthing" are necessary, but soil variations may still obscure the apparent surface patterns. Ground-truthing involves checking (on the ground) certain elements of what was seen by aircraft or satellite.

The experienced mapper relies on the "lay of the land" (perhaps perceptibly slight differences in concavity or convexity) and the quality and vigor of vegetation to help fill in the incomplete observations of substrate conditions. The balance of intensive and extensive work and of "calibration" checks and routine extension of the mapped area may depend greatly on available time, money, and competing duties. Administrative considerations tend to favor statistical control and objective measures of unexplained variance within the somewhat arbitrarily grouped mapping units. Yet for the understanding of vegetation, soil, or other terrain variables, which correlate with the whole ecosystem's condition, nonroutine detectivelike insights are needed, as well as the orderly use of what has been guessed, observed, and confirmed. *See* GLACIATED TERRAIN; SOIL CONSERVATION.

Digital imagery. Digital imagery technologies have revolutionized the way in which vegetation and ecosystem mapping data are collected and analyzed. With the use of sophisticated electronic sensors and computers, workers can extract enormous amounts of information about the Earth's vegetation and ecosystems that was not possible using only aerial photograph analyses. Applications include agricultural crop type mapping and monitoring; forestry clear-cut and burn mapping and species delineation; and land cover and land use mapping of changes in agriculture, urban and rural areas, and biomass. In addition, workers can now measure ecosystem changes that are occurring throughout the world at spatial scales that were never practical with previous technology.

There are two general digital imagery systems: active and passive. Active systems use lasers (for example, laser induced distance and ranging, LIDAR) or radar (for example, synthetic aperture radar, SAR) to send out light or electromagnetic radiation and measure the reflected signal. Passive systems (such as *Landsat Thematic Mapper*, *TM*) measure combinations of wavelengths of electromagnetic radiation reflected differentially from the Earth. *See* LIDAR; RADAR; REMOTE SENSING.

There are three general types of passive digital imagery: panchromatic, multispectral, and hyperspectral. Panchromatic imagery shows vegetation in black and white and has limited use. It sees electromagnetic radiation that is visible to the human eye (for example, *India Remote Satellite*, *IRS*). Multispectral imagery (for example, *Landsat TM*) usually has a dozen or less detection bands, suitable for differentiating gross classes of vegetation. This type of imagery measures the visible spectrum of light (that is, red, green, and blue light) but also adds invisible spectra such as infrared (thermal) and ultraviolet. Hyperspectral imagery, with over 200 bands, can discriminate between individual plant species, or it can identify diseased areas. Most of the spectra measured are not visible to the human eye, so sophisticated computers are used to digest the vast amounts of information. In the future, there will be ultraspectral imagery, containing thousands of bands, increasing the ability to identify as yet unseen properties of the Earth's vegetation. *See* INFRARED RADIATION; ULTRAVIOLET RADIATION.

There are two ways of characterizing the resolution of digital imagery: spectral and spatial. Spectral resolution refers to how many different wavelengths of electromagnetic radiation a sensor can detect. Some hyperspectral sensors (for example, Compact Airborne Spectrographic Imager, CASI) can detect over 250 bands. Spatial resolution refers to the smallest object that can be detected by the sensor. The spatial resolution is usually expressed as a gridcell size. For example, each gridcell of the *Landsat TM* satellite is 30 m (100 ft) on a side. Therefore, anything smaller than 30 m on the ground will not be seen very clearly by the satellite. Some of the newer satellites have sensors that can detect ground features as small as 1 m (39 in.).

Active imagery technologies can gather three-dimensional information about vegetation that is impossible to perceive with passive-type remote sensing. For example, the height and even volume of the tree canopy can be measured. However, the spectral resolution is not as good; that is, species classification is by gross morphology or texture, as opposed to reflected light. Using passive remote sensing, the complex nature of how light of varying wavelengths is reflected from the Earth is measured, so a greater resolution of species types is possible.

Some of the most powerful digital imagery technologies combine active and passive sensors. For example, LIDAR, SAR, and *Landsat TM* sensors can be integrated, giving a unique three-dimensional representation of vegetation species types, morphology, and community structure. In addition, with the LIDAR technology, detailed information about the terrain and elevation is incorporated, facilitating vegetation-elevation correlation analyses.

Map examples. Local detail must be generalized greatly when one attempts an overview of the whole Earth system (see **illus.**).

As a first step, one might focus on environmental characteristics. L. Holdridge's system of life zones, for example, though without seasonal patterns, has been the basis for many maps of potential vegetation, especially in the tropical and subtropical Americas. The resulting data set has been used for predicting the influence of climate change on the world's vegetation patterns.

In addition to vegetation types, various maps have shown estimated world patterns of annual primary production, based both on generalization from measurements and on prediction from climatic relationships. Omission of animal consumption of vegetation may make the estimates low but probably will not change the striking geographic pattern shown. Plant respiration, detrital decomposition, and other basic processes have also been mapped by prediction from climatic relationships.

Structural phenomena, which have more complex environmental relations and trade-offs, have become much easier to model using remote-sensing technology, coupled with direct measurements. The map of the terrestrial storage of carbon by E. Matthews et al. (2000) is based not just on above- and below-ground live vegetation but also on carbon stored in soils. Maps of this nature become powerful tools when they are used to measure changes (natural or from human influence) over time that occur in the Earth's ecosystems. For example, the United Nations Environment Program and the World Conservation Monitoring Center (UNEP-WCMC) created a composite map of the global distribution of original and remaining forests, showing that nearly 50% of the Earth's forests have been destroyed. Details of patterns of ecosystem rates and parameters, on the scale of the world vegetation zones and on the more detailed scale showing the accelerating changes due to humans, continues to be a major challenge for future generations of ecologists.

Numerous international agencies are involved with long-term research on the Earth's ecosystems. Much of their information comes from remote-sensed data for measuring changes in the Earth's vegetation and ecosystems. In addition, further advances in technology have enabled scientists to monitor climate change and predict the impacts of global climate change on the Earth's ecosystems.

Agencies that are responsible for studying global environmental change and its impact on ecosystems include Global Change and Terrestrial Ecosystem (GCTE), Global Terrestrial Observing System (GTOS), and Global Climate Observing System (GCOS). *See* CLIMATE MODELING; GLOBAL CLIMATE CHANGE.

The significance and application of vegetation and ecosystem mapping has changed significantly since the time of Küchler over 50 years ago. The stakes are high, with worldwide deforestation, global climate change, accelerated species extinction rates, and declining biodiversity. Scientists need vast amounts of information about the planet, and they need it quickly in order to save the Earth's various ecosystems. *See* BIODIVERSITY; ECOLOGY; EXTINCTION (BIOLOGY); TERRESTRIAL ECOSYSTEM

. Blake E. Feist; Elgene O. Box

Bibliography. M. P. Anderson et al., *International Classification of Ecological Communities: Terrestrial Vegetation of the United States*, vol. II: *The National Vegetation Classification System: List of Types*, The Nature Conservancy, Arlington, VA, 1998; R. S. Driscoll et al., An Ecological Land Classification Framework for the United States, *U.S. Forest Serv. Misc. Publ.*, no. 1439, 1984; Federal Geographic Data Committee, *Vegetation Classification Standard*, FGDC-STD-005, 1997; D. H. Grossman et al., *International Classification of Ecological Communities: Terrestrial Vegetation of the United States*, vol. I: *The National Vegetation Classification System: Development, Status, and Applications*. The Nature Conservancy, Arlington, VA, 1998; R. Leemans, *Global Data Sets Collected and Compiled by the Biosphere Project*, Working Paper, IIASA-Laxenburg, Austria, 1990; E. Matthews et al., *Pilot Analysis of Global Ecosystems: Forest Ecosystems*, World Resources Institute, 2000; UNESCO, International Classification and Mapping of Vegetation, Series 6 of *Ecology and Conservation*, 1973.

Velocimeter

An instrument that measures the velocity of a flowing liquid or gas, usually understood to be the average velocity of fluid within a relatively small measuring volume. Velocimeters are distinct from flow meters, which measure spatially averaged velocity, or flow rate, across the cross section of a duct or channel. Depending on the nature of the flow and the objective of the measurement, a velocimeter may be used to measure a time-dependent flow velocity or a time-averaged value. Another characteristic of a velocimeter is its spatial resolution, namely the size of its measuring volume. Some velocimeters measure flow velocity in both magnitude and direction, whereas others only resolve a velocity component in a particular direction or plane. The term anemometer is sometimes used as synonymous to velocimeter, although, strictly speaking, the former term should apply only to instruments that measure air velocity. *See* FLOW MEASUREMENT.

Velocimetry is an essential activity in fluid mechanics research laboratories and in flow-related industrial and environmental applications. The operation of different velocimeters is based on a variety of principles:

1. Pressure tubes, including Pitot-static tubes, relate flow velocity to measurable pressure differences in the flow.

2. Thermal anemometers, such as the hot-wire and the hot-film anemometers, relate flow velocity to the rate of convective heat transfer from a heated sensor. Pulsed-wire anemometers track the convection of heat pulses injected into the flow by small sources.

3. Laser Doppler velocimeters relate flow velocity to the frequency shift of laser light scattered by fine particles transported by the flow. The same principle, but applied to sound waves, is utilized by ultrasonic Doppler velocimeters.

4. Particle tracking measures flow velocity by optically monitoring the displacement of particles transported by the flow. A refined and powerful extension of this method is particle image velocimetry.

5. Popular instruments used for monitoring wind speed, such as the cup and propeller anemometers,

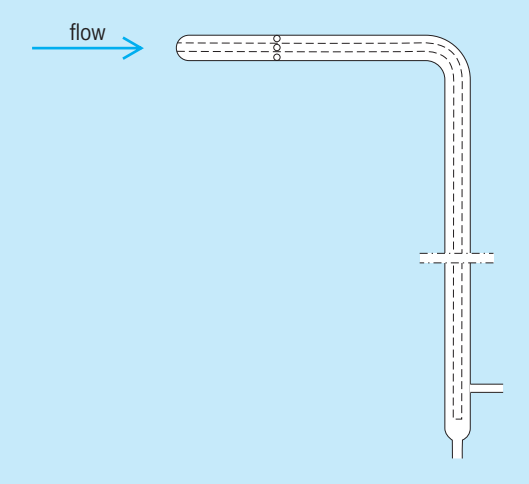


Fig. 1. Pitot-static tube.

measure flow velocity via a force or torque applied on a mechanical element.

Pressure tubes and multihole probes. Pressure tubes are thin hollow tubes inserted into a stream to measure its velocity V, static pressure P, or total pressure P_o. In idealized flow, these three parameters are related to each other through Bernoulli's equation (1), where ρ is the fluid density and the combination $\frac{1}{2}\rho V^2$ is called the dynamic pressure. A Pitot tube, also called total or impact tube, is a simple tube with an open end facing the flow and the other end connected through flexible tubing to a manometer or pressure transducer. A Pitot tube would ideally read the average total pressure across its face. A static tube is also hollow, but, instead, has a sealed, rounded nose facing the flow and a number of small holes on its side. It reads approximately the static pressure in the free flow. A coaxial combination of these tubes, called the Pitot-static tube (**Fig. 1**), reads directly the dynamic pressure as $P_o - P$, from which the flow velocity can be easily calculated with the use of Eq. (1). In this device, the inner tube faces the flow and reads the total pressure, whereas the outer tube is sealed in the nose, has circumferential holes, and reads the static pressure. *See* BERNOULLI'S THEOREM; MANOMETER; PITOT TUBE; PRESSURE MEASUREMENT; PRESSURE TRANSDUCER.

$$P_o = P + \tfrac{1}{2}\rho V^2 \qquad (1)$$

Pressure tubes are simple, inexpensive, and easy to operate as they do not require calibration and their output can be monitored by simple instruments, like a home-made manometer. On the other hand, they also have considerable limitations. Like any other intrusive method, they may distort the flow and its properties, particularly the static pressure. They require alignment within a few degrees with the flow direction, so they cannot be used in flows that keep changing direction or are highly turbulent, such as would be the case with the wind. Their temporal response is slow, typically requiring one minute or more to settle after they are first connected or following a sudden change in flow velocity. Bernoulli's equation is based on the assumptions of steady, incompressible (that is, with constant density) and inviscid (that is, frictionless) flow, so pressure-static tubes would be subjected to errors when used in flows that deviate significantly from these conditions. Examples of inappropriate use include measurements very close to a wall; measurements in fluids with very high viscosity, such as thick oils; measurements in very low speed flows; and when the flow speed is not very small compared to the speed of sound. For air at room temperature, this means that simple pressure tubes are usable only in the range between about 0.3 and 100 m/s (0.7 and 220 mi/h). Compressible flow relationships must be used at higher speeds. *See* COMPRESSIBLE FLOW; VISCOSITY.

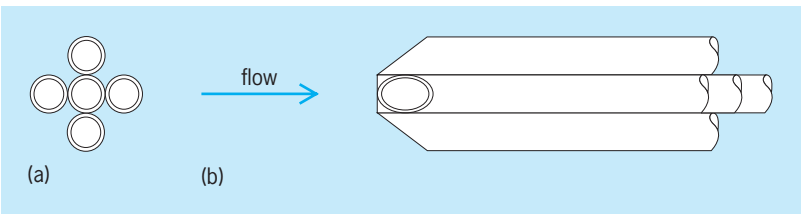


Fig. 2. Five-hole probe. (*a*) Front view. (*b*) Side view.

Different combinations of pressure-measuring devices can be used to measure flow direction (yawmeters) or both flow direction and velocity magnitude. An example is the five-hole probe (**Fig. 2**), which consists of five thin tubes fastened together. The central tube has a square nose facing the flow and measures the total pressure, while the four other tubes have slanted noses and measure pressures that are intermediate between the total and static values, depending on the orientation of the flow velocity. Such devices require precise manufacturing, elaborate calibration over a range of angles, and a sophisticated algorithm to recover the velocity magnitude and direction from the simultaneous readings of the five pressures. *See* YAW INDICATOR.

To overcome the slow response of pressure tubes, researchers have proposed the mounting of miniature pressure transducers within the tubes. Such devices, called fast-response pressure probes, can provide measurements in transient and turbulent flows, but their outputs are also subjected to additional errors due to electronic interference and vibrations.

Thermal anemometers. In thermal anemometry, the sensor is a fine, short metallic element, inserted in the flow and supplied with an electric circuit, such that the generated Joule heat increases its temperature above that of the surrounding flow. For metallic elements, the resistance also rises above its value for the unheated sensor. If the flow velocity rises, then more heat is removed from the sensor, so that its temperature tends to decrease. In the usual thermal anemometry configuration, called constant temperature anemometry (CTA), the sensor comprises one side of a Wheatstone bridge, with the other three sides consisting of resistors. A high-gain difference amplifier senses any imbalance in the bridge as a result of a change in the sensor resistance, and provides

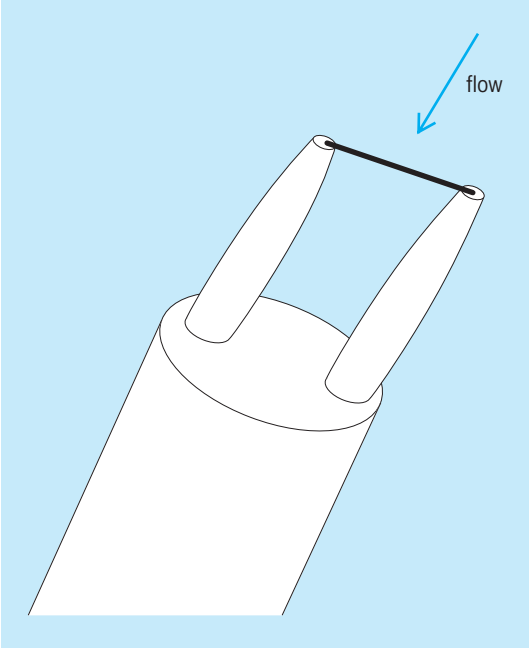


Fig. 3. Hot-wire probe.

additional feedback current that tends to increase the amount of heat generated by the sensor. This tends to increase its temperature, which, in turn, increases its resistance, thus restoring the bridge balance in such a way that the sensor resistance is always kept equal to a preset value. The voltage across the sensor is related to the flow velocity and temperature. With proper calibration, and if the flow temperature is maintained equal to that during calibration or is measured by other means, the flow velocity may be determined from the measured voltage. *See* RESISTANCE MEASUREMENT; WHEATSTONE BRIDGE.

The hot-wire anemometer (HWA) has a cylindrical sensor made of platinum-plated tungsten or platinum alloys. A single-sensor hot-wire probe has an insulated body containing two parallel metallic needles (prongs), at the free tips of which the sensor is mounted and through which current is flowing (**Fig. 3**). A cylindrical sensor is mainly sensitive to the velocity component normal to its axis, and so it cannot resolve any other velocity components. Two- and multisensor hot-wire probes (**Fig. 4**), in which the sensors are inclined with respect to the average flow direction, can, with proper angular calibration, measure two velocity components (cross-wire anemometers) or even the full velocity vector. Hot wires can be used only in gases and nonconducting liquids. Due to their fragility (a typical sensor has a diameter of 5 μm and a length of about 1 mm) and high sensitivity, they require an environment free of dust, droplets, and other particles that may impact on them and break them or alter their heat transfer characteristics. Hot films have sensors that are thin (about 1 μm thick) films of nickel, deposited on a ceramic substrate and coated by a quartz insulation, so that they are insulated from the flowing fluid and can be used in tap water and other conducting liquids. Because of their sturdier construction, they are less sensitive than hot wires to flow impurities and aerodynamic loading.

An algebraic equation for the response of hot wires and hot films may be formulated by considering that Joule heating is the energy input and convective heat transfer is nearly all of the energy output, while the sensor temperature is maintained constant so that there is no change of energy stored in the sensor. This equation connects the sensor voltage E, the flow velocity V, the flow temperature T_f, and and the sensor temperature T_s. Such an equation would also contain geometrical, physical, and electrical properties of the sensor, but such properties cannot be known precisely and also may change as the sensor ages in operation. For this reason, instead of an exact equation, thermal sensor response is described by the semiempirical relationship (King's law) given in Eq. (2), with the coefficients A, B, and n determined

$$\frac{E^2}{T_s - T_f} = A + BV^n \tag{2}$$

by calibration in a flow of known velocity and temperature, such as a small calibration jet. The optimal value of n is near 0.5, but the values of A and B depend on the properties of each individual sensor and the settings of the electrical circuit.

A great advantage of thermal anemometers is their excellent spatial and temporal resolution, making them suitable for statistical measurements in turbulent and unsteady flows. In addition to their fragility and the need for calibration, disadvantages include sensitivity to flow direction, inability to resolve flow orientation in reversing flows, and sensitivity to flow temperature and composition, besides that to flow velocity.

A great variety of thermal anemometer sensors and driving circuits are available commercially at considerable cost. A skilled technician can put together both simple hot-wire probes and electronic circuitry at a fraction of this cost, based on available designs. The recording and processing of thermal anemometer signals is best done with the use of a digital data acquisition system connected to a computer.

Laser Doppler velocimeter. The laser Doppler velocimeter (LDV), alternately referred to as a laser Doppler anemometer (LDA), utilizes the Doppler phenomenon, namely the fact that the frequency of waves scattered by a moving object is different from the frequency of incident waves, with the difference (Doppler frequency) being proportional to

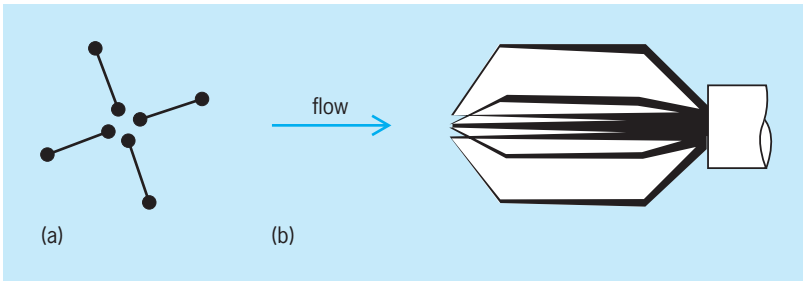


Fig. 4. Four-wire probe. (*a*) Front view. (*b*) Side view.

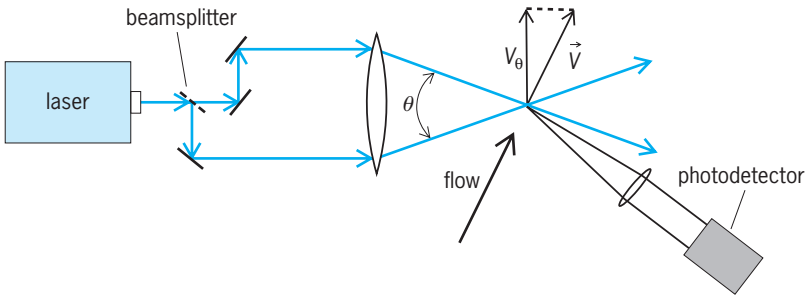


Fig. 5. Laser-Doppler velocimeter.

the component of the velocity of the object along the bisector of the angle between the directions of propagation of the incident and scattered waves. For practical reasons, an optical Doppler system must use a laser beam, which consists of high-intensity, collimated, coherent (that is, with all light waves in phase), monochromatic (that is, with a single frequency, or a small number of distinct frequencies) light. Because the Doppler frequency is far smaller than can be resolved by any available photosensitive device (photodetector, commonly a photomultiplier tube or a photodiode), the light frequency detection system is arranged in a way that it detects the Doppler frequency directly and not as the difference of two slightly unequal frequencies. The most popular configuration is the dual-beam system (**Fig. 5**). In this arrangement, the laser beam is split by a beamsplitter (for example, a partially transparent mirror or prism) and, with the use of additional optical components, forms two parallel beams. These beams are made to intersect in a small measuring volume. A receiving lens collects the light from the measuring volume and projects it on the photodetector, which provides an electric signal that is proportional to the intensity of the received light. *See* DOPPLER EFFECT; LASER.

Because homogeneous fluids, such as clean air and water, would not provide a signal, light-scattering particles are required to cross the measuring volume. Tap water and room air may contain sufficient impurities to act as scatterers, but most commonly it is preferable, or even necessary, to seed the flow, that is, to introduce a small amount of solid particles or liquid droplets in an even distribution. Compared to natural impurities, flow seeding has the advantage of allowing control of the size and shape of the particles, thus reducing errors that would be caused by particles of different masses moving at different speeds relative to the fluid. Clearly, it is important that the motion of foreign particles matches the motion of the surrounding fluid, which makes it necessary to use very small particles, typically of dimensions comparable to 1 μm.

In a dual-beam system, the velocity component V_θ normal to the bisector of the angle θ between the intersecting beams is determined from the frequency f_D of the collected light to be given by Eq. (3),

$$V_\theta = \frac{\lambda}{2\sin(\theta/2)} f_D \qquad (3)$$

where λ is the wavelength of the laser light. Thus, it is only necessary to measure f_D and no calibration is required. Among the various methods for determining the frequency of the photodetector signal, the one most widely used currently is burst analysis. In this method, it is ensured that the particle concentration is sufficiently small for no more than a single particle to reside in the measuring volume at any time instant and a dedicated device computes the peak of frequency spectrum of the light emitted by a single particle (burst) by performing a fast Fourier transform (FFT). *See* FOURIER SERIES AND TRANSFORMS.

Compared to hot-wire anemometry, laser Doppler velocimetry has the advantage of being nonintrusive and usable in media containing impurities. As it measures true local velocity, it is insensitive to fluid temperature and composition. As with hot wires, the laser Doppler velocimeter system described above cannot discriminate flow direction, but this capability can be easily realized through a method called frequency shifting. In terms of frequency response and spatial resolution, laser Doppler velocimetry is inferior to hot-wire anemometry.

To measure more than one velocity component, it is necessary to use light of different frequencies. A popular approach is to separate the light of an argon-ion laser into two pairs of green and blue beams on two perpendicular planes; both pairs are made to intersect within the same measuring volume, the two colors are separated by optical filters in the receiving system, and each single-color component is processed by a separate photodetector and burst analyzer.

Laser Doppler velocimeter systems are available commercially from several suppliers and are too complex for the common user to put together from generic components. Their cost is quite high, but less-expensive educational systems are also available.

Particle image velocimeter. Particle tracking is the direct method of determining flow velocity by timing the displacement of markers that move with the flow. In the simplest case of readily visible, slowly moving, isolated particles, it is sufficient to simply employ a ruler and a stopwatch, but for the method to be applied to complex flows with high resolution, it is usually necessary to capture consecutive particle positions in rapid succession and to be able to separate and identify the images of closely spaced particles. This need has led to the development of digital particle image velocimetry (DPIV or PIV). In a basic particle image velocimetry system (**Fig. 6***a*), a laser beam is converted to a sheet of light with a small thickness by passing through a glass rod (or special lenses for improved quality). The flow is seeded with small particles (typically 1 μm or less in diameter) and two images of a small portion of the light sheet are recorded, illuminated by two light pulses. Although light pulses can be produced by using a shutter or a rotating disk to interrupt the beam of a continuous-wave laser, such as an argon-ion laser, the most powerful method is to use a twin-cavity Nd:YAG (neodymium yttrium aluminium garnet) laser, which

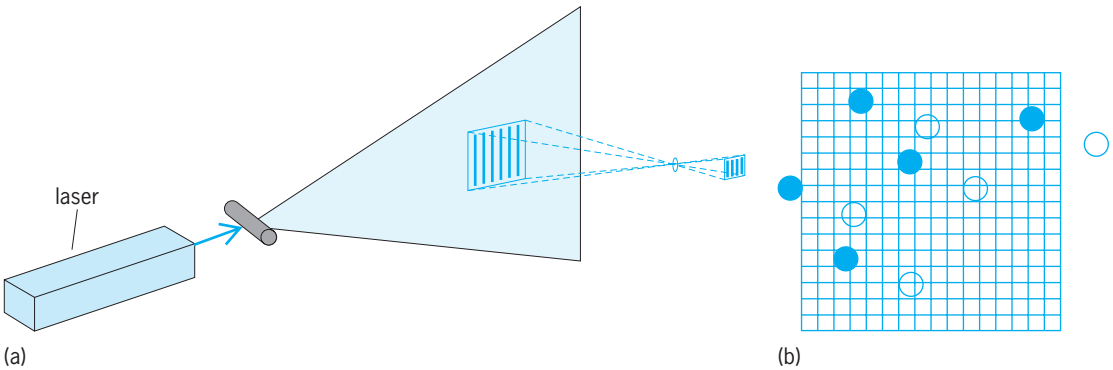


Fig. 6. Particle image velocimeter. (*a*) Apparatus. (*b*) Superposition of two images with pairs of images of particles that are within the measuring volume.

produces high-energy (100–400 mJ) light pulses separated by a time difference δt between 100 ps and 10 ns. Digital photography has advanced sufficiently to allow the independent recording of the two images on two different digital frames. When the particle concentration is sufficiently large, the superposition of the two frames contains a large number of pairs of images of the different particles that are within the measuring volume (Fig. 6*b*). These images can be sorted statistically with the use of sophisticated algorithms employing the cross-correlation technique. Then, each particle's velocity projection on the light sheet plane can be computed by dividing the particle displacement by δt.

Whereas hot-wire anemometry and laser-Doppler velocimetry measure an average flow velocity within the measuring volume, thus being point-measuring methods, particle image velocimetry provides the velocity distribution within an area. This is achieved by dividing the recorded image into smaller elements, called interrogation areas and each containing several particle images, and then determining the velocity in the center of each area by averaging the velocities of the corresponding particles. This makes it possible to draw instantaneous streamlines (that is, lines tangential to the velocity vector) in the flow and compute additional properties, such as the vorticity component normal to the image plane. For this reason, particle image velocimetry has gained great popularity. As laser, camera, and software technology keeps improving, it is expected that the power of particle image velocimetry will further increase. A variation of the basic method is stereoscopic particle image velocimetry, which uses two cameras at different angles to resolve the full three-dimensional velocity vector. Another improvement is time-resolving particle image velocimetry, which provides pairs of particle images at a succession of relatively closely spaced times, thus improving the temporal resolution of the basic method, which is much lower than those of hot-wire anemometry and laser Doppler velocimetry. Early particle image velocimetry systems were put together by individual research teams, but a variety of commercial systems are currently available. As with many other electronic devices, particle image velocimetry systems have improved in quality, while also becoming more affordable.

Ultrasonic velocimeters. The operation of the ultrasonic Doppler velocimeter is based on the same physical principle as laser Doppler velocimetry, but utilizes ultrasound rather than light. The sound is produced by a transmitter, which also alternates as a receiver, collecting scattered sound as it returns toward it. The average velocity of particles scattering sound along the path of the beam is determined from the difference in the frequencies of the transmitted and received signals. By measuring the time delay between a transmitted and a received sound pulse, it also becomes possible to determine the location of the particle, from which it is possible to obtain not only the average velocity, but also an instantaneous velocity profile along the beam path. A disadvantage of this method is that it cannot measure directly the axial velocity component in pipe and duct flows. A great advantage, compared to laser Doppler velocimetry and particle image velocimetry, is that it does not require optical accessibility, and therefore it is suitable for opaque fluids.

Another kind of ultrasonic velocimeter, used primarily in atmospheric research, has a pair of transmitter-receiver transducers facing each other. It measures the average flow velocity along the sound path from the difference in the speed of sound propagating in the two directions along this path, considering that sound travels faster if it propagates with the flow than against it. *See* SOUND; WIND MEASUREMENT.

Cup and propeller anemometers. The cup anemometer (**Fig. 7**) is a familiar device, frequently installed on towers near airports and in meteorological stations for measuring wind speed. It consists of three or four hollow hemispherical cups mounted at some radial distance from a rotating shaft and such that all the concave sides face in the same circumferential direction. When the instrument is exposed to a flowing stream, it will always rotate in the same sense, because the drag on the concave side of each cup is higher than the drag on its convex side. With its shaft vertical, the cup anemometer would rotate at an angular velocity ω which is proportional to the magnitude V_h of horizontal wind speed, but it is

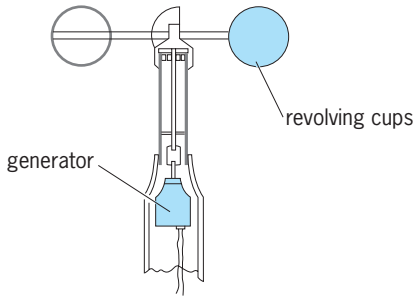


Fig. 7. Cutaway diagram of a rotating cup anemometer. (*After G. K. McMillan and D. M. Considine, eds., Process/Industrial Instruments and Controls Handbook, 5th ed., McGraw-Hill, 1999*)

incapable of resolving wind direction or the vertical wind component. For this reason, it is often accompanied by a rotating vane, which aligns with the wind direction. To measure the vertical wind magnitude, a second cup anemometer can be used, this time with its shaft horizontal. The speed of rotation is monitored by an electric generator coupled to the shaft. All that is required is the proportionality coefficient between ω and V_b, which is obtained by calibration. Although the measuring volume of cup anemometers is relatively large (typically 200 mm or 8 in. across), it is acceptable for most atmospheric applications. In gusty and strongly turbulent winds, cup anemometers tend to overspeed, that is, to read wind speeds higher than the actual one. Moreover, their accuracy deteriorates at speeds lower than 0.25 m/s (0.5 mi/h).

A related device is the propeller anemometer, which actually operates like a wind turbine. It consists of a small axial propeller mounted on a rotating shaft, whose speed of rotation is proportional to the flow velocity parallel to its axis. A commonly used device is the Gill anemometer, which has four helicoidal blades. Like cup anemometers, propeller anemometers require calibration to determine the proportionality coefficient. They also tend to overspeed in turbulent winds, but much less than cup anemometers. If not aligned with the flow direction, they tend to read values lower than the actual wind velocity projection on their axis. Three propeller anemometers mounted on mutually orthogonal shafts can be used to measure wind in both magnitude and direction. Stavros Tavoularis

Bibliography. H. H. Bruun, *Hot-Wire Anemometry*, Oxford University Press, 1995; E. O. Doebelin, *Measurement Systems: Application and Design*, 5th ed., McGraw-Hill, 2004; R. J. Goldstein (ed.), *Fluid Mechanics Measurements*, 2d ed., Taylor & Francis, 1996; F. Mayinger and O. Feldman (eds.), *Optical Measurements: Techniques and Applications*, 2d ed., Springer, 2001; M. Raffel, C. E. Willert, and J. Kompenhans, *Particle Image Velocimetry: A Practical Guide*, Springer, 1998; S. Tavoularis, *Measurement in Fluid Mechanics*, Cambridge University Press, 2005; J. G. Webster (ed.), *Mechanical Variables Measurement: Solid, Fluid, and Thermal*, CRC Press, 2000.

Velocity

The time rate of change of position of a body in a particular direction. Linear velocity is velocity along a straight line, and its magnitude is commonly measured in such units as meters per second (m/s), feet per second (ft/s), and miles per hour (mi/h). Since both a magnitude and a direction are implied in a measurement of velocity, velocity is a directed or vector quantity, and to specify a given velocity completely, the direction must always be given. The magnitude only is called the speed. *See* SPEED.

Linear velocity. A body need not move in a straight line path to possess linear velocity. The instantaneous velocity of any point of a body undergoing circular motion is a vector quantity, such as v_1 or v_2 in **Fig. 1**. When a body is constrained to move along a curved path (**Fig. 2**), it possesses at any point an

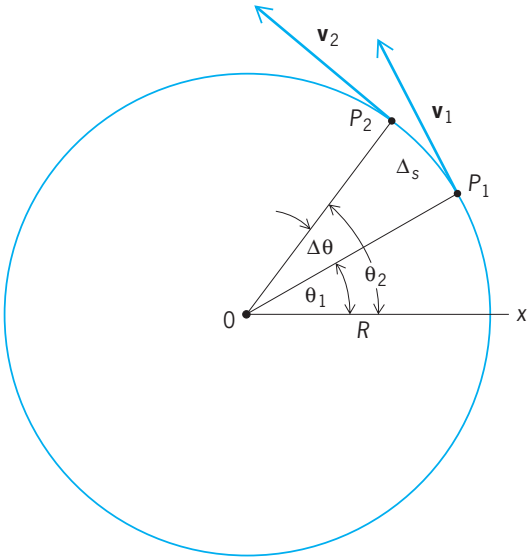


Fig. 1. Illustration of angular displacement, angular speed, and tangential velocity.

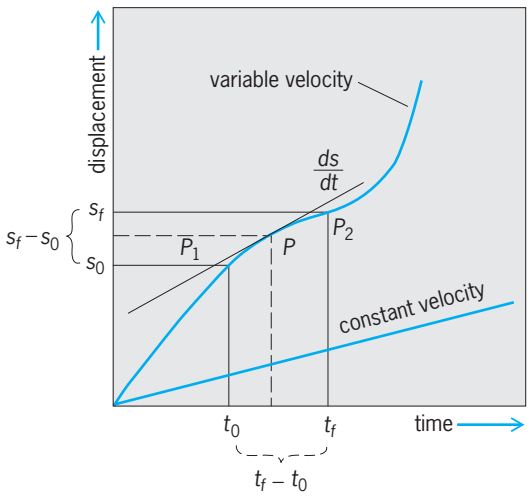


Fig. 2. Average velocity from P_1 to P_2 is $(s_f - s_0)/(t_f - t_0)$. The instantaneous velocity at point P is the limit of the ratio representing the average velocity as the interval approaches zero.

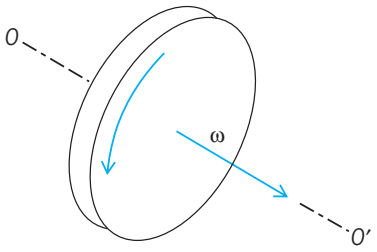


Fig. 3. Angular velocity shown as an axial vector.

instantaneous linear velocity in the direction of the tangent to the curve at that point. The average value of the linear velocity is defined as the ratio of the displacement to the elapsed time interval during which the displacement took place. The displacement of a body from an initial position s_0 to a final position s_f after time t is equal to $s_f - s_0$. The corresponding time interval is $t_f - t_0$. The magnitude of the average velocity is then given by Eq. (1), where Δs is

$$\bar{v} = \frac{\text{displacement}}{\text{elapsed time}} = \frac{s_f - s_0}{t_f - t_0} = \frac{\Delta s}{\Delta t} \tag{1}$$

displacement and Δt is the corresponding elapsed time.

The magnitude of the instantaneous velocity v of a body is the limiting value of the foregoing ratio as the interval approaches zero. In the notation of calculus, Eq. (2), ds/dt is the instantaneous time rate of change

$$v = \lim_{\Delta t \to 0} \frac{\Delta s}{\Delta t} = \frac{ds}{dt} \tag{2}$$

of displacement (Fig. 2).

The velocity of a body, like its position, can only be specified relative to a particular frame of reference. Consequently, all velocities are relative. *See* RELATIVE MOTION.

Angular velocity. The representation of angular velocity ω as a vector is shown in **Fig. 3**. The vector is taken along the axis of spin. Its length is proportional to the angular speed and its direction is that in which a right-hand screw would move. If a body rotates simultaneously about two or more rectangular axes, the resultant angular velocity is the vector sum of the individual angular velocities. Thus, if a body rotates about an x axis with angular velocity ω_x, and simultaneously about a y axis with an angular velocity ω_y, the resultant angular velocity ω is the vector sum given by Eq. (3).

$$\omega = \omega_x + \omega_y \tag{3}$$

It should be emphasized that whereas angular velocities are commutative in addition, that is, they may be added in any order, angular displacements are not commutative. *See* ROTATIONAL MOTION.

Angular displacement. Figure 1 represents a body rotating with circular motion about an axis through O perpendicular to the figure. Line OP_1 is the position of some radius in the body at a time t_1, with θ_1 being the angular displacement from a reference line. Line OP_2 is the position of the same radius at a later time t_2, with the angular displacement θ_2. Angular displacement may be measured in degrees, radians, or revolutions.

Angular speed. From Fig. 1, it is seen that the body has rotated through the angle $\Delta\theta = \theta_2 - \theta_1$ in the time $\Delta t = t_2 - t_1$. The average angular speed $\overline{\omega}$ is defined by $\overline{\omega} = \Delta\theta/\Delta t$, and the instantaneous angular speed $\omega = d\theta/dt$. Although it is customary in most scientific work to express angular speed in radians per second, it is common in engineering practice to use the units of revolutions per minute (rpm) or revolutions per second (rps).

Tangential velocity. When a particle rotates in a circular path of radius R through an angular distance $\Delta\theta$ in a time Δt, as in Fig. 1, it traverses a linear distance Δs. The average linear speed $\overline{v}$ is given by Eq. (4), since $\Delta s = R\Delta\theta$. Similarly, the instantaneous

$$\bar{v} = \frac{\Delta s}{\Delta t} = \frac{R\ \Delta\theta}{\Delta t} = \bar{\omega} R \tag{4}$$

speed v is given by $v = \omega R$. The direction of this instantaneous speed is tangential to the circular path at the point in question. Any vector **v** drawn in this direction represents the tangential velocity.

Combined velocities. A body may have combined linear and angular motions, as is the case when the wheel of a moving automobile rolls along the ground with an angular velocity about its axle which moves with a linear velocity parallel to the pavement. In this case, a point on the rim of the tire describes a curved path called a cycloid. If a circular body rolls on the surface of a sphere, a point on the periphery of the rotating body describes a curve called an epicycloid. *See* CYCLOID; EPICYCLOID.

Rogers D. Rusk

Bibliography. J. D. Cutnell and K. W. Johnson, *Physics*, 4th ed., 1997; H. Goldstein, *Classical Mechanics*, 3d ed., 2001; D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics*, 6th ed., 2001; C. Kittel, W. D. Knight, and M. A. Ruderman, *Mechanics*, vol. 1, 2d ed., 1973.

Velocity analysis

A technique for the determination of the velocities of the parts of a machine or mechanism. Both graphical and analytical analyses of plane mechanisms will be discussed in this article, but of the several methods of each type available, only one of each type will be described. The graphical method will be discussed first, since the visualization which is an inherent part of the graphical analysis generally gives a better physical feel for the problem than most purely analytical methods. Analytical methods, however, are necessary for computer analyses. *See* VELOCITY.

Purposes. In a high-speed machine it is important that the inertia forces be determined. This requires an acceleration analysis of the machine, and the first step in an acceleration analysis usually is a velocity analysis. In the analysis of some machines, the velocity of a particular point in the machine may itself be the important thing to be determined in the analysis—for example, the cutting speed and return

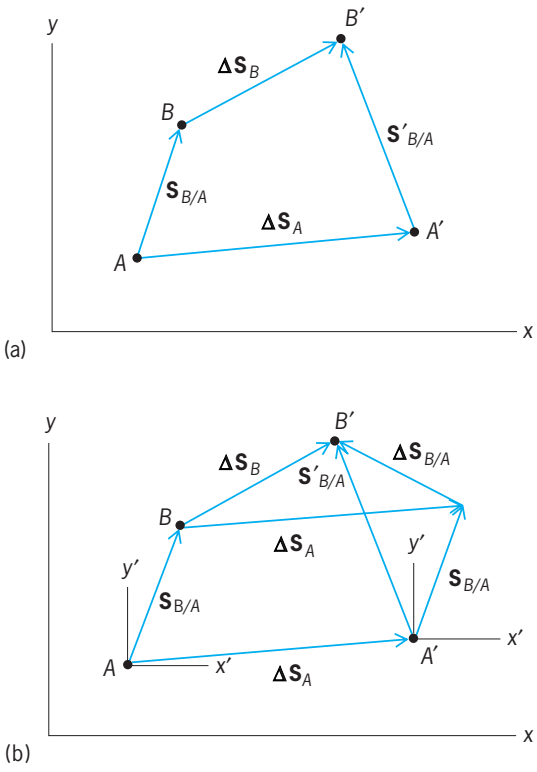


Fig. 1. Relative displacement. (*a*) Position vectors locating *B* with respect to *A* at beginning and end of a time increment. (*b*) Vectors for derivation of relative displacement equation.

speed of the cutting tool in a shaper, or the shuttle velocity in a textile machine.

Relative displacement. The method of relative velocities is widely used in velocity analysis, and the relative velocity equation is fundamental for this method of analysis. It is derived by first considering relative displacements. Referring to **Fig. 1***a*, let A and B be any two points moving in the stationary x-y coordinate system shown. Point B is located with respect to point A by the position vector $\mathbf{S}_{B/A}$. Suppose that during a time interval Δt, point B has the vector displacement $\Delta\mathbf{S}_B$ and point A the displacement $\Delta\mathbf{S}_A$. At the end of the time interval Δt, B is located with respect to A by the vector $\mathbf{S}'_{B/A}$. The vector change in $\mathbf{S}_{B/A}$ during the time interval Δt is shown in Fig. 1*b* as $\Delta\mathbf{S}_{B/A}$. From Fig. 1*b* it is evident that Eq. (1) holds, which is the relative displacement

$$\Delta\mathbf{S}_B = \Delta\mathbf{S}_A + \Delta\mathbf{S}_{B/A} \tag{1}$$

equation. It is further noted from Fig. 1*b* that $\Delta\mathbf{S}_{B/A}$ is the displacement of B as measured in the nonrotating x'-y' coordinate system which is attached to and moves with A.

Relative velocity. If the time interval Δt approaches zero, the displacements shown in Fig. 1*b* become infinitesimal and the velocity vectors representing the velocities of the points are proportional to the infinitesimal displacement vectors as shown in **Fig. 2**. From Fig. 2 the relative velocity equation (2) can

$$\mathbf{v}_B = \mathbf{v}_A + \mathbf{v}_{B/A} \tag{2}$$

be written. Equation (2) may also be derived mathematically by dividing Eq. (1) through by Δt and taking the limit as Δt approaches zero, as shown by Eq. (3) or (4). It must be clearly understood that $\mathbf{v}_{B/A}$

$$\lim_{\Delta t\to 0}\frac{\Delta\mathbf{S}_B}{\Delta t} = \lim_{\Delta t\to 0}\frac{\Delta\mathbf{S}_A}{\Delta t} + \lim_{\Delta t\to 0}\frac{\Delta\mathbf{S}_{B/A}}{\Delta t} \tag{3}$$

$$\mathbf{v}_B = \mathbf{v}_A + \mathbf{v}_{B/A} \tag{4}$$

in Eq. (2) represents the velocity of B as measured in a nonrotating coordinate system attached to A. It is usually called the velocity of B relative to A, although there has been some objection to this terminology.

Two points on a rigid body. Let body 2 of **Fig. 3** represent any mechanism link moving with general plane motion. Its angular velocity is ω (the vector representation of ω is a vector out of the paper). Equation (2) holds for any two points, and therefore holds for points A and B of the special case being considered. The relative velocity $\mathbf{v}_{B/A}$ is the velocity of B measured in the nonrotating x'-y' coordinate system attached to A. In this coordinate system B moves on a circular path with its center at A, and therefore its velocity as measured in this coordinate system has a magnitude equal to $r_{AB}\omega$ and a direction perpendicular to line AB. *See* RIGID-BODY DYNAMICS.

Coincident points on two links. Consider the mechanism of **Fig. 4***a* in which the pin at the end of link 2 is constrained to move in the slot of 3. Point P_2 is the center of the pin, and point P_3 is the point on link 3 coincident with P_2. There is no such actual physical point, but an extension built onto link 3 and overlapping

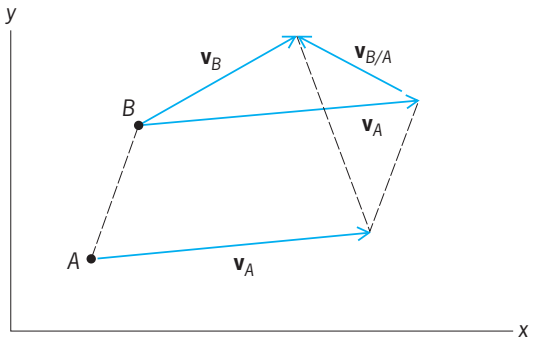


Fig. 2. Vectors for derivation of the relative velocity equation.

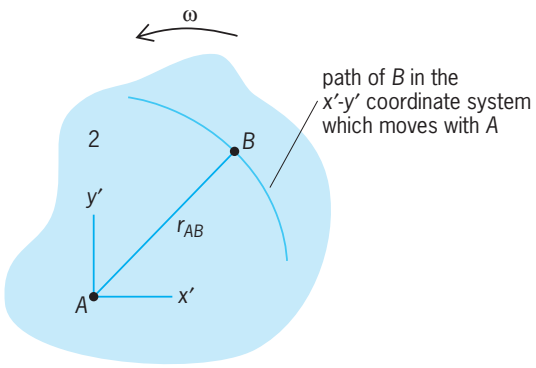


Fig. 3. Rigid body 2 moving with general plane motion.

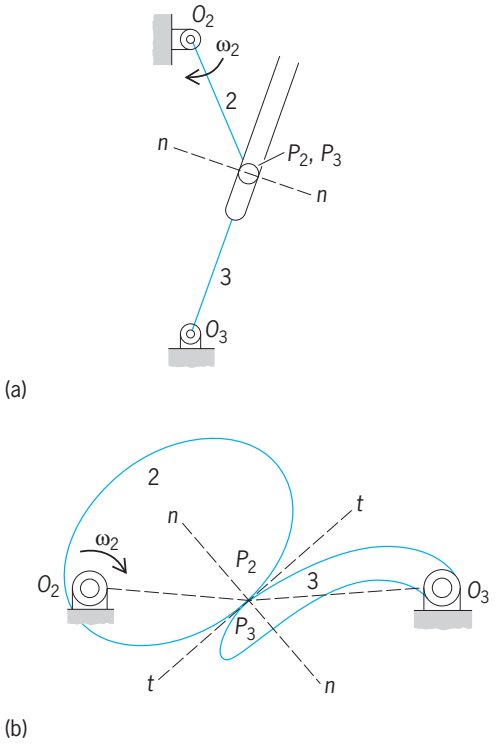


Fig. 4. Mechanisms *a* and *b* for which the relative velocity equation is applied to coincident points on different links.

the pin can be imagined. Equation (2) as applied to the coincident points P_2 and P_3 is Eq. (5). Assume

$$\mathbf{v}_{P_3} = \mathbf{v}_{P_2} + \mathbf{v}_{P_3/P_2} \tag{5}$$

the magnitude of $\mathbf{v}_{P2}$ is known. The directions of $\mathbf{v}_{P2}$ and $\mathbf{v}_{P3}$ are known. To solve a single-vector equation, there can be no more than two unknowns, so that the direction of $\mathbf{v}_{P3}/P_2$ must also be known. There can be no relative motion of points P_2 and P_3 in the direction of line *n-n* normal to the slot, so that the only relative motion possible is along the slot. Therefore the direction of $\mathbf{v}_{P3}/P_2$ is along the slot. Similarly for the direct-contact mechanism of Fig. 4*b*, the relative velocity $\mathbf{v}_{P3}/P_2$ must be along the common tangent of the contacting surfaces, line *t-t'*. *See* LINKAGE (MECHANISM).

Application example. The use of the relative velocity equation will be illustrated in the velocity analysis of a slider-crank mechanism. Slider-crank mechanisms are used in a wide variety of machines. Perhaps the most familiar example is the crank, connecting rod, and piston mechanism of the internal combustion engine. A skeleton drawing of such a mechanism is shown in **Fig. 5***a*. The moving links are numbered 2, 3, and 4, with link 1 being the fixed frame. Key points are identified with letters. It is assumed that the angular velocity of crank 2 is known. The magnitude of the velocity of point A is given by $v_A = r_{O_2A}\omega_2$. A vector representing this velocity is shown in Fig. 5*a* originating at point A, and also is laid off to scale from an origin O_v in Fig. 5*b*. The relative velocity equation, $\mathbf{v}_B = \mathbf{v}_A + \mathbf{v}_{B/A}$, is next applied to points A and B. Point B is constrained to move along a horizontal straight line so that the direction of $\mathbf{v}_B$ is known. Vector $\mathbf{v}_A$ is completely known; $\mathbf{v}_{B/A}$ is perpendicular to line AB, since A and B are two points on the same rigid body (line 3). A velocity polygon (triangle) can be drawn as in Fig. 5*b* to solve for the unknown quantities of the relative velocity equation, the magnitudes of $\mathbf{v}_B$ and $\mathbf{v}_{B/A}$. These magnitudes can be scaled directly from the polygon. Usually it would be necessary to make an analysis of this type for a number of positions of the mechanism in its motion cycle. Other mechanisms can be analyzed graphically for velocities in a similar manner. *See* SLIDER-CRANK MECHANISM.

Use of vector mathematics. In the graphical method of velocity analysis just discussed, the geometry of the mechanism is known in each phase of its motion cycle from the drawing of the mechanism in each position. In an analytical velocity analysis, the geometry is usually determined analytically; that is, a position analysis of the mechanism is performed by using trigonometry, vector mathematics, or some other analytical method as the first step in the velocity analysis.

Position analysis. One analytical method for the position analysis of mechanisms, developed by M. A. Chace, makes use of vector mathematics. It consists essentially of solving vector triangles containing two unknowns. For example, in **Fig. 6** an offset slider-crank mechanism is shown with appropriate vectors placed on the mechanism. Assuming the position of crank 2 is known as the input to the mechanism, vectors **P** and **Q** are known and may be added together to obtain vector **C**. Vectors **C**, **r**, and **s** form a vector triangle containing two unknowns, the direction of vector **r** and the magnitude of vector **s**. That is, unit vector $\hat{r}$ and s are the unknowns (carets are used in this article to designate unit vectors). The plane vector equation $\mathbf{C} + \mathbf{s} + \mathbf{r} = 0$ (or $\mathbf{C} + s\hat{s} + r\hat{r} = 0$) must be solved for the unknowns. Chace

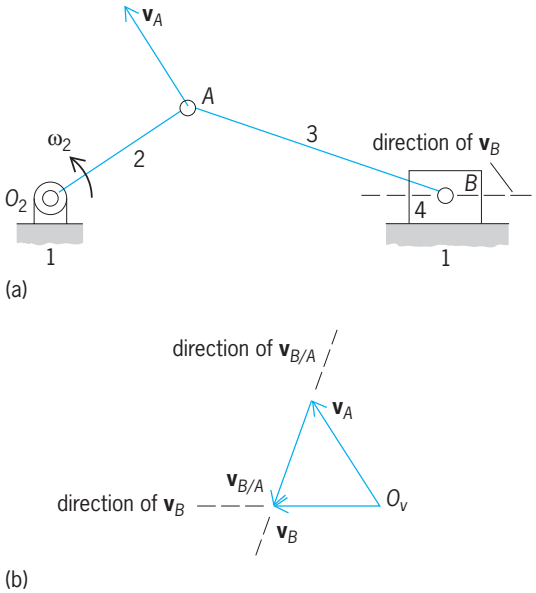


Fig. 5. Velocity analysis. (*a*) Slider-crank mechanism to be analyzed. (*b*) Velocity polygon.

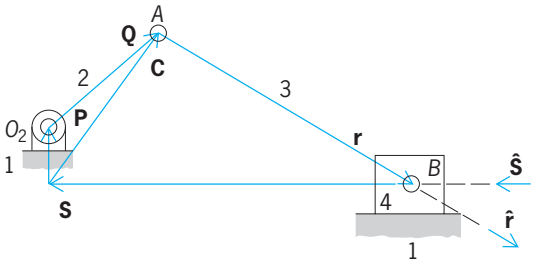


Fig. 6. Offset slider-crank mechanism showing vector triangle to be solved for position analysis of the mechanism.

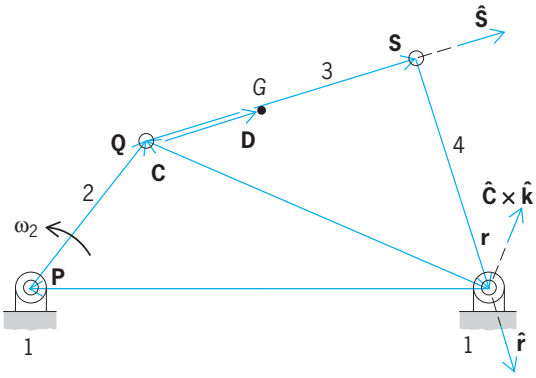


Fig. 7. Four-bar linkage showing vector triangle to be solved for the position analysis of the mechanism.

developed solutions for the plane vector equation for various cases using vector mathematics. Vector **C** is assumed always to be the known vector, and case 1 is the situation where two magnitudes, s and r, are unknown. Chace's equations for the vectors **s** and **r** for this case are Eqs. (6) and (7), where $\hat{k}$ is the unit

$$\mathbf{s} = \frac{\mathbf{C} \cdot (\hat{\mathbf{r}} \times \hat{\mathbf{k}})}{\hat{\mathbf{r}} \times (\hat{\mathbf{s}} \times \hat{\mathbf{k}})} \hat{\mathbf{s}} \tag{6}$$

$$\mathbf{r} = \frac{\mathbf{C} \cdot (\hat{\mathbf{s}} \times \hat{\mathbf{k}})}{\hat{\mathbf{s}} \cdot (\hat{\mathbf{r}} \times \hat{\mathbf{k}})} \hat{\mathbf{r}} \tag{7}$$

vector in the z direction or out of the paper.

Case 2 is the situation illustrated in Fig. 6 where there is one unknown magnitude, s, and one unknown direction, $\hat{\mathbf{r}}$, in the vector triangle. Chace's equations for the vectors **r** and **s** for this case are Eqs. (8) and (9). The lower signs on the radicals

$$\mathbf{r} = -[\mathbf{C} \cdot (\hat{\mathbf{s}} \times \hat{\mathbf{k}})](\hat{\mathbf{s}} \times \hat{\mathbf{k}}) \pm \sqrt{r^2 - [\mathbf{C} \cdot (\hat{\mathbf{s}} \times \hat{\mathbf{k}})]^2}\hat{\mathbf{s}} \tag{8}$$

$$\mathbf{s} = \left(-\mathbf{C} \cdot \hat{\mathbf{s}} \mp \sqrt{r^2 - [\mathbf{C} \cdot (\hat{\mathbf{s}} \times \hat{\mathbf{k}})]^2}\right)\hat{\mathbf{s}} \tag{9}$$

would be chosen for the analysis of this mechanism because **r** has a negative component in the $\hat{s}$ direction for the complete mechanism cycle.

Case 3 is the situation illustrated in **Fig. 7**. Vector **P** connecting the fixed pivots is a known vector and can be expressed in the form $\mathbf{P} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}}$. The position of the input crank 2 is assumed to be known, so that vector **Q** is known and can be expressed in the same form as vector **P**. Vectors **P** and **Q** can be added together to get vector **C**. Vectors **C**, **s**, and **r** form a vector triangle in which the unknowns are unit vectors $\hat{\mathbf{s}}$ and $\hat{\mathbf{r}}$. Chace's solutions for vectors **r** and **s** for this case are Eqs. (10) and (11). The upper

$$\mathbf{r} = \mp\sqrt{r^2 - \left(\frac{r^2 - s^2 + C^2}{2C}\right)^2}(\hat{\mathbf{C}} \times \hat{\mathbf{k}}) - \left(\frac{r^2 - s^2 + C^2}{2C}\right)\hat{\mathbf{C}} \tag{10}$$

$$\mathbf{s} = \pm\sqrt{r^2 - \left(\frac{r^2 - s^2 + C^2}{2C}\right)^2}(\hat{\mathbf{C}} \times \hat{\mathbf{k}}) + \left(\frac{r^2 - s^2 + C^2}{2C} - C\right)\hat{\mathbf{C}} \tag{11}$$

signs on the radicals would be selected for the analysis of the mechanism of Fig. 7, because **s** has a positive component in the $\hat{\mathbf{C}} \times \hat{\mathbf{k}}$ direction and **r** has a negative component in that direction. *See* FOUR-BAR LINKAGE.

Time derivative of a vector. The time derivative of a vector originating from a fixed origin such as the one shown in **Fig. 8** is the velocity of the point at the tip of the vector; that is, Eq. (12) holds.

$$\mathbf{v}_P = \lim_{\Delta t \to 0} \frac{\Delta \mathbf{r}}{\Delta t} = \frac{d\mathbf{r}}{dt} = \dot{\mathbf{r}} \tag{12}$$

Time derivative of a fixed-length vector. If a vector **r** in a rotating body is of fixed length and originates on the axis of rotation of the body as shown in **Fig. 9**, its time derivative is $\omega \times \mathbf{r}$, since the magnitude of $\omega \times \mathbf{r}$ is $\omega r \sin\phi$, and it is evident from Fig. 9 that the velocity of point P (the time derivative of **r**) is $\omega r \sin\phi$ in the direction of $\omega \times \mathbf{r}$. It may also be shown that the time derivative of any fixed-length vector **r** in a body is equal to $\omega \times \mathbf{r}$ even though the vector **r** does not originate on the axis of rotation of the body. In this case, however, the derivative of the vector is not the velocity of the point at the tip of the vector, but the velocity of the point at the tip of the vector relative to the point at the origin or tail of the vector.

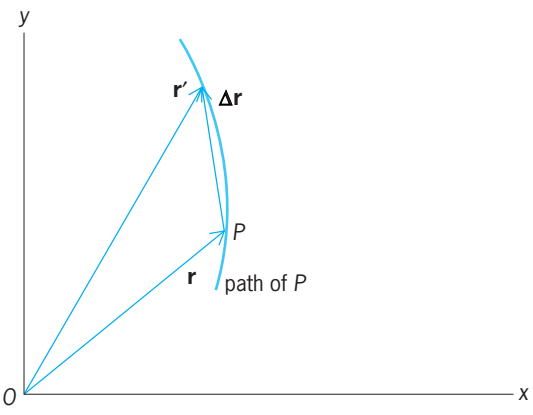


Fig. 8. Changing vector originating from a fixed origin.

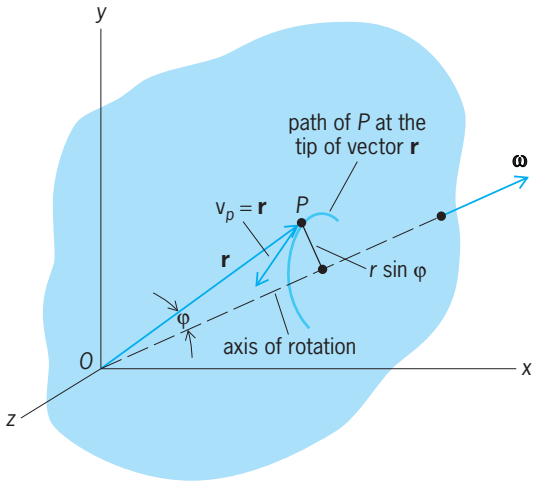


Fig. 9. Fixed-length vector in a body which originates on the axis of rotation of the body; $v_p = r = \omega \times r$.

Four-bar mechanism. For the four-bar mechanism shown in Fig. 7, the vector loop equation (13) may

$$\mathbf{P} + \mathbf{Q} + \mathbf{s} + \mathbf{r} = 0 \qquad (13)$$

be written. Differentiating Eq. (13) with respect to time yields Eq. (14), since vectors **Q**, **s**, and **r** are all fixed-length vectors. In a plane motion mechanism the directions of ω_3 and ω_4 are known (direction $\hat{\mathbf{k}}$). The unknowns ω_3 and ω_4 may be determined by taking the dot product of each vector of Eq. (14) with

$$\omega_2 \times Q + \omega_3 \times s + \omega_4 \times r = 0 \qquad (14)$$

$\hat{\mathbf{r}}$ and $\hat{\mathbf{s}}$, respectively, reducing Eq. (14) in each case to a scalar equation with a single unknown. The solutions thus determined for ω_3 and ω_4 are Eqs. (15) and (16). With these angular velocities known, the

$$\omega_3 = \frac{-\omega_2(\hat{\mathbf{k}} \times \mathbf{Q}) \cdot \hat{\mathbf{r}}}{s(\hat{\mathbf{k}} \times \hat{\mathbf{s}}) \cdot \hat{\mathbf{r}}} \qquad (15)$$

$$\omega_4 = \frac{-\omega_2(\hat{\mathbf{k}} \times \mathbf{Q}) \cdot \hat{\mathbf{s}}}{r(\hat{\mathbf{k}} \times \hat{\mathbf{r}}) \cdot \hat{\mathbf{s}}} \qquad (16)$$

velocity of any point on the mechanism can be found. For example, the velocity of point G is given by Eq. (17).

$$\mathbf{v}_G = \omega_2 \times Q + \omega_3 \times D \qquad (17)$$

Computer subprograms for Chace solutions to the plane vector equation and for all the various vector operations have been written, thus making the position and velocity analyses (as well as the acceleration analysis) of mechanisms by computer using Chace's method quite convenient. Computer analyses are particularly useful when complex mechanisms must be analyzed in many positions of their cycle of motion. *See* STRAIGHT-LINE MECHANISM; VECTOR (MATHEMATICS).

James C. Wolford

Bibliography. S. Doughty, *Mechanics of Machines*, 1988; B. Paul, *Kinematics and Dynamics of Planar Machinery*, 1979; J. Shigley and C. R. Mischke, *Machine Design Fundamentals*, 1989; J. Shigley and J. Uicker, *Theory of Machines and Mechanisms*, 2d ed., 1994.

Veneer

A thin sheet of wood of uniform thickness produced by peeling, slicing, or sawing. Depending on the manner of production and the portion of wood from which a veneer is made, the grain may be flat, vertical, or biased. Most veneer is rotary-cut from a bolt of wood, called a flitch, centered in the chucks of a lathe, which may have a capacity for logs from 2 to 16 ft (0.6 to 4.8 m) long, about 8 ft (2.4 m) being a common capacity. A nose bar bears against the flitch parallel to the center line of the lathe and a knife, also extending nearly the length of the lathe, peels off the veneer. Knives near the ends of the flitch cut the edges of the veneer.

The peeling blade advances to maintain the thickness of the veneer at the initial setting, which may be as thin as $^1/_{64}$ in. (0.4 mm), more typically in the range from $^1/_{32}$ to $^1/_{20}$ in. (0.75 to 1.25 mm), or as thick as $^3/_8$ in. (9.5 mm). The outer surface of the veneer as it emerges from under the nose bar is smooth; it is called the tight side. However, the inner surface tends to form lathe checks as the blade curls it away from the log; hence it is called the loose side.

The veneer is clipped to approximate size, dried sufficiently to prevent attack from molds and fungi (moisture typically 15% or less by weight), and bundled until needed. Each bundle contains the sheets cut from a single flitch, stacked in order as cut, and labeled as to total content. The user can then match veneers for figured panels.

Veneers cut from selected hardwoods, burls, crotches, and stumps are used for facings on furniture and the interior decoration of buildings. The art of veneering was practiced in Egypt at least as early as the fifteenth century B.C. Until the twentieth century most veneer was produced by sawing; for the hardest woods, such as ebony and oak, veneer is still sawed. During the eighteenth century veneering (in which a single relatively large veneer covers an area) and marquetry (in which numerous small pieces of wood and other materials form a pattern) reached a peak of craftsmanship in the decoration of fine furniture. During the early years of the industrial revolution, ornate inlays and highly figured veneers applied on poorly fabricated furniture created an impression that veneering was a cheap substitute for solid construction. With improved methods of drying and gluing, veneering regained its position in furniture manufacture and has been extended to wall paneling and other decorative purposes. Other uses of veneer include plate separators in storage batteries, boxes for fruit and vegetables, drums for cheeses, and crates, hampers, and baskets for transportation and storage. Spanish cedar, sliced to $^1/_{100}$ in. (0.25 mm), is used to wrap fine cigars. Most veneer is used in plywood panels. *See* PLYWOOD; WOOD PRODUCTS.

Frank H. Rockett

Ventilation

The supplying of air motion in a space by circulation or by moving air through the space. Ventilation may be produced by any combination of natural or mechanical supply and exhaust. Such systems may include partial treatment such as heating, humidity control, filtering, or purification, and, in some cases, evaporative cooling. More complete treatment of the air is generally called air conditioning. *See* AIR CONDITIONING.

Natural ventilation. Natural ventilation may be provided by wind force, convection, or a combination of the two. Although largely supplanted by mechanical ventilation and air conditioning, natural ventilation still is widely used in homes, schools, and commercial and industrial buildings. It is effective and economical in areas of prevailing winds and for high industrial buildings that have hot equipment which provides the motivating convective force.

Wind-force ventilation may be provided by direct force, such as wind blowing in one side of a building and out the other side or through a monitor. This is commonly known as cross ventilation. Because of the friction and velocity losses incurred through building openings, only about 25–60% of the wind velocity is available for ventilation, depending upon whether the wind direction is perpendicular or oblique to the building openings. The resulting negative pressures on the downwind side of the building (if open to the outside) and the sizes and locations of the exhaust openings all materially affect the flow of air.

Airflow around an object creates a negative pressure on the downwind side. This principle is used to advantage in the design of a large number of building ventilators and monitors to provide basic ventilation, to assist existing convective ventilation, or to prevent backdraft down into a building. **Figure 1** illustrates the round, rotating head, and continuous roof-type ventilators. Because of the many complex forces involved, airflow capacities cannot be calculated but must be determined by testing.

The force for convection ventilation is created by the difference in weight of two air columns at different temperatures; the heavy, cool column attempts to displace the hot, light column. The pressure p exerted by a fluid column varies as the height h and density w of the fluid; that is, $p = hw$, expressed in consistent units. The basic equation for such flow is $V = \sqrt{2gh}$, again in consistent units for velocity V, height h, and gravity g. From these relations and because temperature has a direct relationship to density, the equation below may be used to estimate con-

$$Q = 9.4A\sqrt{h(t_i - t_o)}$$

vective flow (**Fig. 2**). Here Q = air passing through an opening by convective force, ft³/min; A = free (net) area of the opening, ft² (the inlet areas are assumed to equal the exhaust openings); h = height from inlets to outlets, ft; t_i = average temperature of indoor air column in height h, °F; t_o = temperature of outdoor air, °F; and 9.4 = constant of proportionality, including value of 65% for effectiveness of openings.

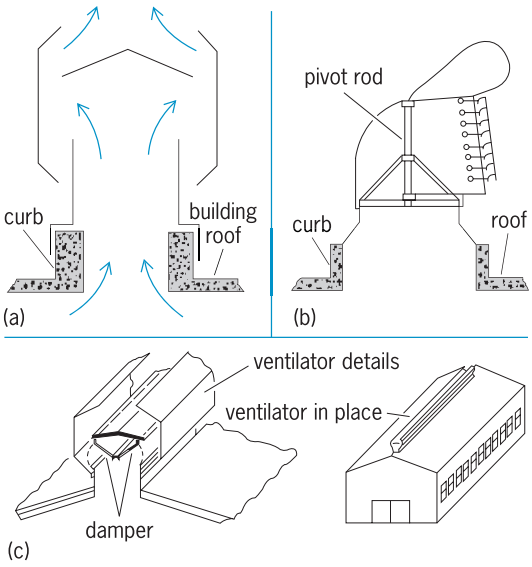


Fig. 1. Roof exhausts for natural ventilation. (*a*) Cross section of round ventilator and (*b*) of rotating-head ventilator. (*c*) Continuous roof ventilator.

This equation indicates that convective velocities tend to be low. Consequently, high buildings having large exhaust (and inlet) openings and high internal heat loads are the most practical application for convection ventilation. The best flow would be obtained by a well-shaped hole in the roof, but, because weather protection must also be provided, it is customary to exhaust through roof monitor or baffle-type exhaust ventilators as in Figs. 1 and 2 for small volumes and **Fig. 3** for large volumes.

Mechanical ventilation. Mechanical supply ventilation may be of the central type consisting of a central fan system with distributing ducts serving a large space or a number of spaces, or of the unitary type (Fig. 3) with little or no ductwork, serving a single space or a portion of a large space. Both types are employed for schools and for commercial and industrial applications. *See* DUCTED FAN.

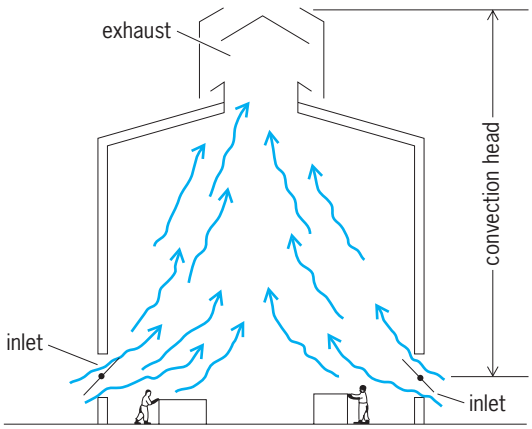


Fig. 2. Building cross section illustrating motivating convection head (height) for natural ventilation.

Central system assemblies may be custom-built, factory-prefabricated for job site assembly, or factory-assembled. Ventilating units are factory-assembled with rare exceptions. The assemblies consist of a fan and usually include air filters and air heaters of the finned-tube type. The fans may be of the propeller, axial flow, or centrifugal type. Central systems require the last two fan types because of the higher static pressures usually encountered in the distributing ducts. *See* FAN.

The mechanically powered roof ventilator is a practical source of low-cost ventilation. For summer (nontempered) supply, such units consist of fans and weather hoods, or dampers (**Fig. 4**). For all-year ventilation, heaters or other equipment may be added.

Outside air connections are generally provided for all systems. Outside air is needed in controlled quantities to remove odors and to replace air exhausted from the various building spaces and equipment. The inlets are located to minimize the intake of fumes, dust, organic materials, and pollens. Because it is never possible to find completely clean air, ventilation air filters are usually provided in the system casings in all except a few industrial or summer relief applications. This prevents clogging and poor heat transfer for the air heaters and helps to reduce the pollen and dust in the occupied areas served by the systems.

Air distribution. Duct systems to distribute and disperse the air are required for all but the small unitary systems to avoid short-circuiting and to provide adequate ventilation to all parts of the space served. Such distribution permits desirable air movement in the space without undesirable draft and temperature stratification. Ventilation systems generally serve as conveyors for adding or removing heat and humidity. The distribution system must be adequate for this purpose also. The amount of air circulated is important because too small a volume requires uncomfortably high temperatures for heating or results in high building temperature for heat removal ventilation. Similar problems occur with humidity because of the limited amounts of moisture which can be conveyed without condensation problems. If a system is oversized, it becomes unnecessarily expensive. In addition, it may be difficult to discharge the air in the space within acceptable velocity limits.

Outlet grilles and diffusers of the rectangular, square, and round type are provided to further control and distribute the supply of air within the selected throw (blow length) for each outlet. Any air column acts as a pump and will entrain many times the primary volume of air. The outlets are designed and selected to obtain maximum entrainment (and mixing) because this greatly reduces air motion and temperature stratification, making possible greater comfort for the room occupants.

Exhaust ventilation systems. Exhaust ventilation is required to remove odors, fumes, dust, and heat from an enclosed occupied space. Such exhaust may be of the natural variety previously described or may be mechanical by means of roof or wall exhaust fans or mechanical exhaust systems. The mechanical

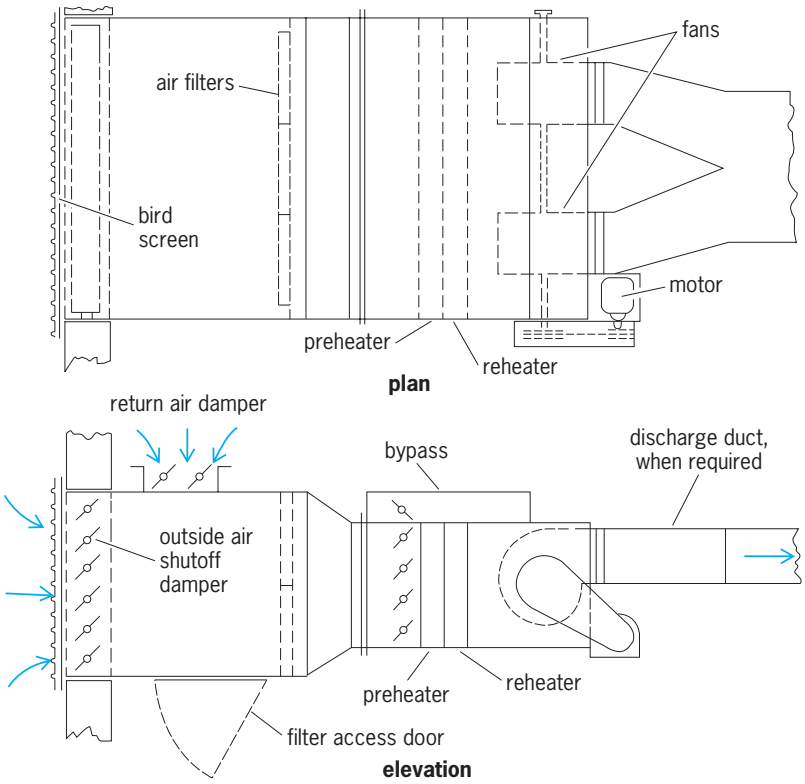


Fig. 3. Factory-assembled ventilating unit.

systems may have minimal ductwork or none at all, or may be provided with extensive ductwork which is used to collect localized hot air, gases, fumes, or dust from process operations. Where it is possible to do so, the process operations are enclosed or hooded to provide maximum collection efficiency with the minimum requirement of exhaust air.

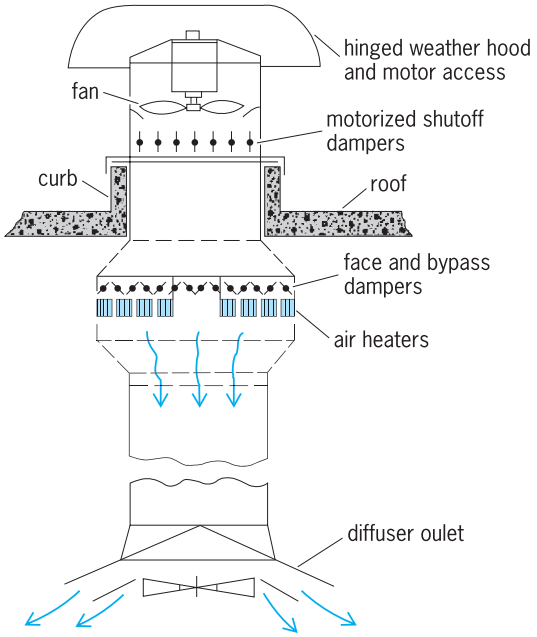


Fig. 4. Mechanically powered roof air supply unit. For summer ventilation, the portion shown in solid lines is used. Heater and dampers are for winter use. Diffuser outlet improves air distribution into ventilated space.

Because of the possibilities of recirculated or external air pollution, it is customary to remove dust and fumes where practical or where required by means of ordinary ventilation filters, more efficient washers or centrifugal collectors, or chemical scrubbers.

Where dust is conveyed in the exhaust ducts, the velocities must be adequate to lift and move the dust particles. Velocities of 3000–6000 ft/min (914–1828 m/min) are required for this purpose. Lower velocities are tenable for fume removal, but corrosion protection must be provided by selection of duct and equipment materials. The round duct is used in most dust and fume systems because of its lower friction and because of its better dust-handling characteristics.

Fans for ordinary ventilation exhaust or heat removal may be similar to supply fans. Fume and dust exhaust fans must be more rugged and are frequently of the radial-blade (paddle-wheel) type for this reason. Axial flow and conventional centrifugal fans are also used in these applications. John H. Clarke

Venus

The second planet in distance from the Sun. Venus has been called "Earth's twin" because it is similar to Earth in gross characteristics such as mass, radius, and density (see **table**). In other ways Venus is apparently different. Its atmospheric mass is almost a hundred times that of the Earth; its atmosphere is mostly carbon dioxide instead of nitrogen and oxygen; an extensive cloud layer of concentrated sulfuric acid is present; its surface temperature is an unbearable 867°F (464°C); and it rotates with a period of 243 days, and from east to west, clockwise as we look down from the north, in the opposite sense of most other planets. Some of these differences are due more to alternate evolutionary paths of the two planets than to totally different initial conditions.

Appearance. To the naked eye, Venus is the brightest starlike object in the sky. It is usually visible during the night either soon after sunset or close to sunrise. It can sometimes be seen during the daytime. As observed through a pair of binoculars or a telescope, Venus exhibits a crescentlike or gibbous appearance. Similar to the Moon, Venus ranges over a full set of phases from a "new moon" to a "full moon." This phase variation is caused by a changing fraction of the Sun-illuminated hemisphere facing toward the Earth. At new moon, only the dark nighttime side is seen, while at full moon all of the daytime hemisphere is seen. Venus undergoes a complete set of phase changes over its synodic period of 584 days. *See* PHASE (ASTRONOMY).

Clouds. The light seen coming from Venus is entirely due to sunlight that is reflected from a dense cloud layer whose top is located about 45 mi (70 km) above the surface and whose bottom lies within 30 mi (50 km) of the surface. In contrast to the Earth's approximately 50% cloud cover, the clouds of Venus are present over the entire planet. In yellow or red light they present a uniform appearance. However, the clouds show a banded and spotted pattern when viewed in ultraviolet light (**Fig. 1**). These ultraviolet markings provide information about atmospheric motions.

The clouds of Venus consist of a large number of tiny particles, about 1 micrometer in size, that are made of a water solution of concentrated sulfuric acid. Such a composition may at first seem very surprising when compared with the water clouds of the Earth's lower atmosphere. However, sulfuric acid particles are the dominant type of particles in the Earth's upper atmosphere, although the amount there is much less than the amount present in Venus's atmosphere. In the case of both planets, sulfuric acid is produced primarily from sulfur-containing gases that combine with water vapor and oxygen-containing gases. Compositional measurements made from the United States *Pioneer Venus* Sounder probe, which descended through Venus's atmosphere on December 9, 1978, show that sulfur dioxide is the principal sulfur-containing gas in Venus's atmosphere. Sulfur dioxide is also the major gas species injected into the Earth's stratosphere by volcanic explosions. Such injections cause a large, but temporary, increase in the amount of sulfuric acid there.

Atmospheric composition. By far, the chief gas species of Venus's atmosphere is carbon dioxide, which makes up 96% of the atmospheric molecules, while nitrogen accounts for almost all the remainder. Trace amounts of sulfur dioxide (150 parts per million), water vapor (20 ppm), carbon monoxide (17 ppm), argon (70 ppm), helium (12 ppm), neon (7 ppm), hydrogen chloride (0.4 ppm), and hydrogen fluoride (0.005 ppm) are present in the lower atmosphere, with the concentration of the first two of these declining dramatically near the cloud tops due to the formation of new sulfuric acid there. Chemical transformations also occur in the deeper portions of the atmosphere, aided by the high temperatures there. For example, at altitudes below about 25 mi (40 km), carbon monoxide is gradually converted into carbonyl sulfide, a gas containing carbon, oxygen, and sulfur atoms.

Carbon dioxide. In contrast to the dominance of carbon dioxide in Venus's atmosphere. The Earth's atmosphere consists mostly of nitrogen and oxygen, with carbon dioxide being present at a level of only 340 ppm. In part, this difference may stem more from temperature differences than from intrinsic differences. Over the lifetime of the Earth, an amount of carbon dioxide comparable to that in Venus's atmosphere was vented out of the Earth's hot interior. The outgassed carbon dioxide remained in the atmosphere for only a short time. Almost all of it participated with rain in dissolving land rocks. Rivers carried the dissolved rock and carbon dioxide into the oceans, where they subsequently precipitated to form carbonate rocks, such as limestone. Venus's surface is much too hot for oceans of water to be present, and hence its atmosphere has been able to retain essentially all of the carbon dioxide vented from its interior. *See* ATMOSPHERIC CHEMISTRY.

Characteristics of Venus*

Characteristics	Values
Mass	0.815 Earth's mass
Equatorial radius	0.95 Earth's equatorial radius or 3760.4 mi (6051.8 km)
Mean density	0.95 Earth's value or 5.24 g/cm^3
Range of topography	9 mi (15 km) [Earth is 12 mi (20 km)]
Gravity	0.907 Earth's gravity
Mean distance from Sun	0.723 Earth's distance
Perihelion	66,780,000 mi (107,480,000 km), 0.718 Earth's distance
Aphelion	67,690,000 mi (108,940,000 km), 0.728 Earth's distance
Orbital period	0.615 of an Earth year or 224.7 Earth days
Orbital eccentricity	0.4 Earth's value or 0.0068
Orbital inclination to Earth's orbital plane	3.39°
Sidereal rotational period	243.686 Earth days with respect to the stars
Length of day	116.75 Earth days with respect to the Sun
Atmospheric surface pressure	92 bars (1334 lb/in.^{2}), 90 times Earth's atmospheric pressure
Average surface temperature	867°F (464°C) [Earth is 59°F (15°C)]
Surface wind speed	0.7–2.2 mi/h (0.3–1.0 m/s)
Surface atmospheric composition (by volume)	96.5% carbon dioxide (CO_2), 3.5% nitrogen (N_2); minor constituents (ppm): sulfur dioxide (SO_2): 150; argon (Ar): 70; water (H_2O): 20; carbon monoxide (CO): 17; helium (He): 12; neon (Ne): 7

*Some of these numbers depend on assumptions as to when and where they were observed and some variation will be seen in the literature.

Rare gases. There are two varieties of rare gases found in planetary atmospheres: those that were derived from the gas cloud (solar nebula) from which the planets formed (primitive rare gases), and those that were produced from the radioactive decay of certain elements, such as potassium, in the interior of the planets. A fundamental finding about the composition of Venus's atmosphere is the detection of much more primitive argon and neon than in Earth's atmosphere. Furthermore, Mars's atmosphere has an even smaller amount of these rare gases than does the Earth's. These differences in the abundances of rare gases among the atmospheres of Venus, Earth, and Mars may, in part, be due to sizable differences in the rates at which their earliest atmospheres were lost. Such losses may have been caused by the blowoff of portions of these atmospheres by massive, high-velocity stray bodies that were particularly abundant in the first several hundred million years of these planets' lifetimes. The blowoff was powered by hot rock vapor generated when these bodies collided with the planets' surfaces. Mars was especially vulnerable to atmospheric blowoff because of its low mass and hence its low surface gravity. Conceivably, the Earth lost a good fraction of its earliest atmosphere when a body with a mass comparable to that of Mars hit it. Such a giant impact is considered by many as the most likely way by which the Earth's Moon formed. Since Venus and Earth have comparable amounts of carbon dioxide and nitrogen in their atmospheres and the rocks of their interiors, these atmospheric components may have been added at a later stage, perhaps from volatile-rich bodies that came from the outer solar system. *See* EARTH; MARS; MOON; SOLAR SYSTEM.

Water vapor. The amount of water vapor in Venus's atmosphere is much less (about 100,000 times) than the amount of water in the Earth's oceans. Since it is unlikely that Venus was initially endowed with so much less water than the Earth, Venus probably lost almost all of its original water over its lifetime. The loss of water from a planet may be determined primarily by the amount that is in the upper atmosphere, the stratosphere, where solar ultraviolet radiation decomposes water vapor molecules into hydrogen and oxygen. The light gas hydrogen can eventually escape the planet's gravity and be lost to space, while the leftover oxygen can combine with other gases, such as carbon monoxide, or with iron at the planet's surface. Because Venus is closer to the Sun than the Earth, its lower atmosphere was hotter,

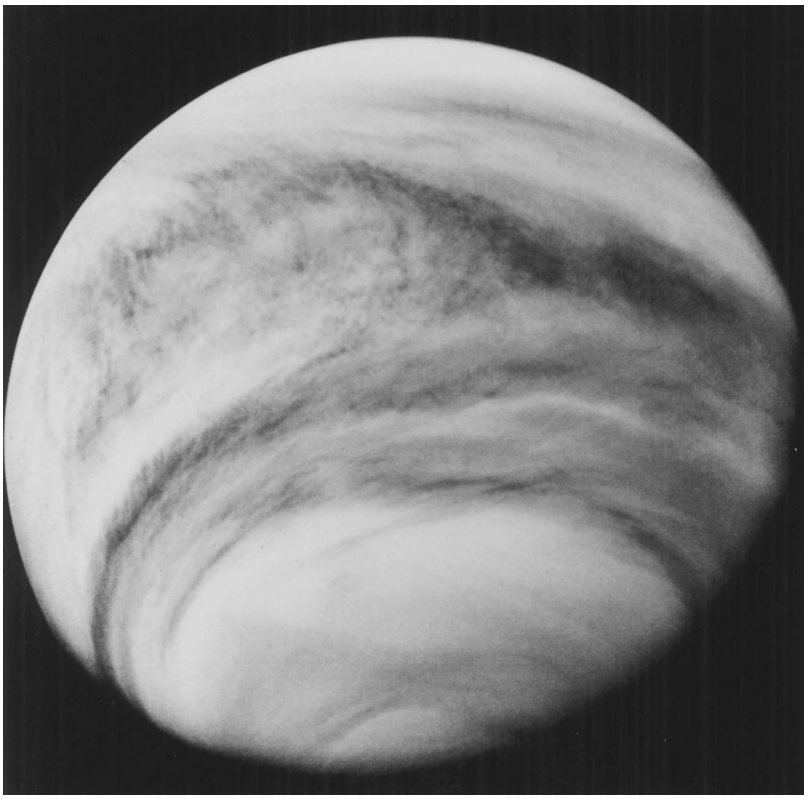


Fig. 1. Clouds of Venus photographed in ultraviolet light from the *Pioneer Venus* spacecraft. The south pole is near the bottom of the image.

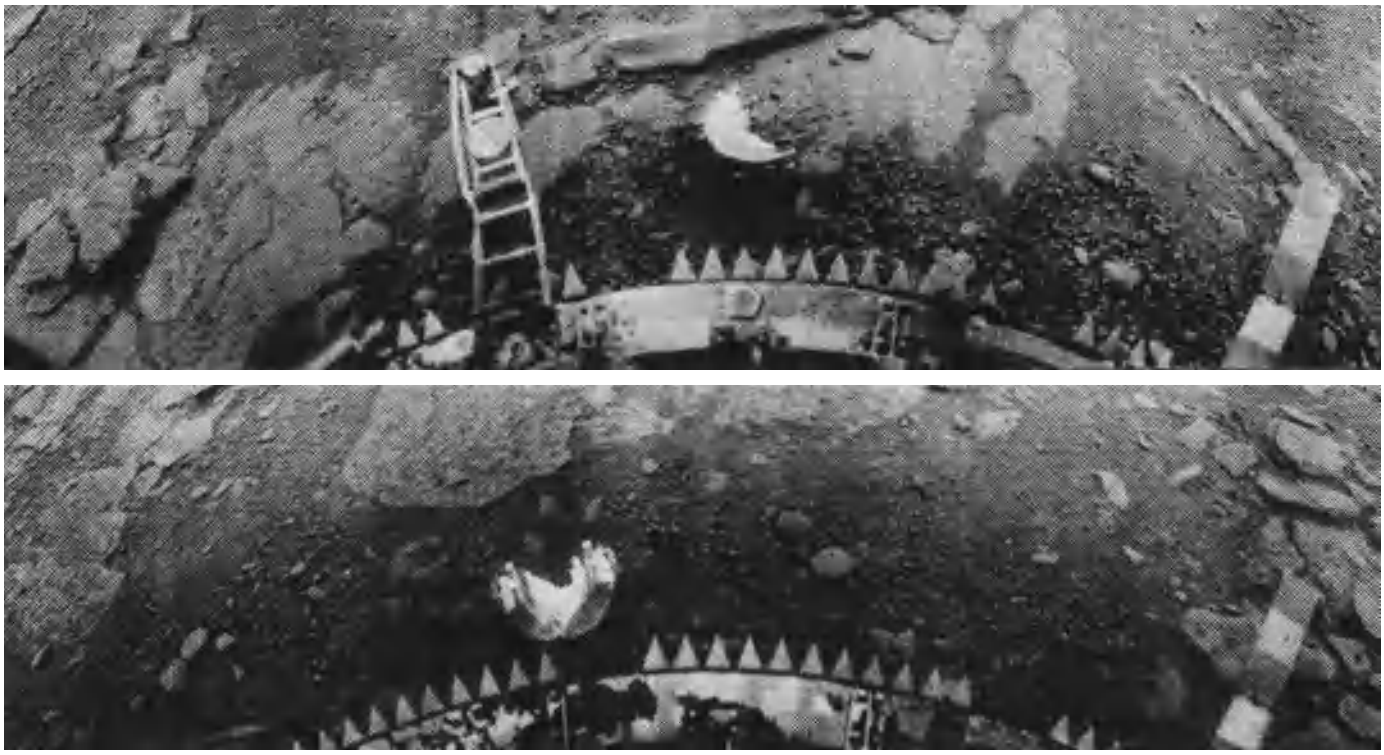


Fig. 2. Panoramic view, in two frames, of the surface of Venus obtained by the Soviet *Venera 13* spacecraft shortly after it landed. Parts of the spacecraft and apparatus extending from it can be seen at the bottoms of the frames.

water vapor was more abundant at early times there, and hence much more of the water vapor was able to penetrate into its stratosphere than was the case for the Earth. Consequently, Venus could have lost much more water over its lifetime than did the Earth. Some confirmation of this viewpoint has been given by the finding of one of the. *Pioneer Venus* experiments that there is about 100 times as much heavy water, or water containing deuterium, on Venus as on Earth. Such an enrichment of deuterium is expected since light hydrogen can more easily escape to space than heavy hydrogen.

Temperature. By detecting long-wavelength heat radiation produced at the surface, radio telescopes first showed that the surface temperature was 850°F (730 K), that is, over 600°F (350 K) higher than the boiling point of water. This result was confirmed by direct temperature measurements of the atmosphere made from Soviet and United States spacecraft that descended through Venus's atmosphere. The atmospheric temperature has a relatively cool value of −10°F (250 K), about 45°F (25 K) below the freezing point of water, near the top of the cloud layer, which is at a pressure of about 1/20 that at the Earth's surface. The temperature gradually increases with decreasing altitude until it reaches about 850°F (730 K) at the surface, where the pressure is 90 times that at the Earth's surface. Just as on Earth, surface temperatures on Venus decrease at higher elevations. At the top of Maxwell Montes, about 7.5 mi (12 km) higher than the average elevation, the temperature could be only about 650°F (620 K).

The high value of Venus's surface temperature is not due to its being closer to the Sun than the Earth. Beacuse its cloud layer reflects to space about 75% of the incident sunlight, Venus actually absorbs less solar energy than does the Earth. Rather, the high temperature is the result of a very efficient greenhouse effect that allows a small but significant fraction of the incident sunlight to penetrate to the surface (about 2.5% according to Soviet and United States spacecraft measurements), but prevents all except a negligible fraction of the heat generated by the surface from escaping directly to space. The thermal energy produced by the surface and hot lower atmosphere, which occurs at infrared wavelengths, is very effectively absorbed by the carbon dioxide, water vapor, sulfur dioxide, and sulfuric acid particles of the atmosphere. However, these materials are poor absorbers at visible wavelengths, where most of the solar energy lies. The greenhouse effect raises Venus's surface temperature by almost 900°F (500 K), but causes only a modest 60°F (35 K) rise in the surface temperature of the Earth. This difference is due to the Earth's atmosphere being partially transparent at some infrared wavelengths, thus permitting some surface heat to escape to space. *See* GREENHOUSE EFFECT.

Meteorology. The atmosphere near the cloud tops is moving with a jet-stream-like velocity of about 225 mi/h (100 m/s) from east to west in the direction of Venus's rotation. Although the solid surface rotates around in 243 days (clockwise as one looks down from the north, opposite the rotation of Earth), the clouds are observed to move around the planet

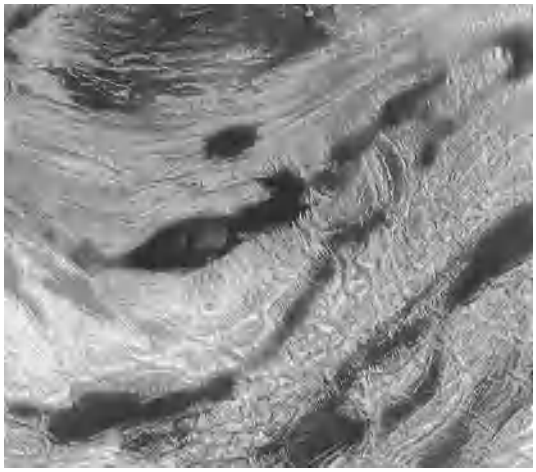


Fig. 3. ***Magellan*** **image of the central and northern parts of the highland region Ovda Regio (north at top). Ridges running approximately east-west (horizontally) have been interpreted as resulting from crustal shortening (compression). Ridges are cut by through-going fractures and graben (fault-bounded valleys), indicating that an episode of northeast-southwest extension followed the compressional event. The youngest event was flooding of low-lying areas by smooth (radar-dark) lava flows. (*Jet Propulsion Laboratory*; *NASA*)**

in the same direction with a period of about 4 days. This motion is called superrotation. Only Saturn's moon Titan shares this phenomenon with Venus. There is a small component of cloud motion of about 10 mi/h (5 m/s) from the equator toward the pole. In contrast to the situation for the Earth where only a small portion of the atmosphere moves at jet-stream speeds, the entire cloud top region on Venus moves at these large speeds. The wind speed for the most part gradually, but in a few places sharply, decreases with declining altitude and achieves values of a modest few meters per second within 6 mi (10 km) of the surface. Winds at the surface are also a few meters per second. However, wind streaks are seen on the surface, in the *Magellan* radar images, in the vicinity of the more recent impact craters, indicating that the formation of these craters produced temporary stormlike atmospheric winds near the surface. *See* SATURN.

Like the winds on the Earth, the winds on Venus are produced ultimately by differences in the amount of solar energy absorbed by different areas of the planet, such as areas at different latitudes. The large wind speeds near the cloud tops may be the result of Venus's atmosphere being very deep. The transport of momentum by the mean circulation and the eddies from the deep dense portions near the surface to the much less dense regions near the cloud tops translates sluggish motions into fast ones. At midlatitudes and in the polar regions of the Earth, the east-west wind speeds are determined primarily by a balance between variations of pressure with latitude and the Coriolis force due to the rotation of the Earth's surface, which is shared by its atmosphere. But Venus's surface rotates too slowly for the Coriolis force to be important in its atmosphere. Rather, centrifugal forces (analogous to those on a merry-go-round) owing to the rotating motions of the winds themselves around the planet may provide the balance to the pressure gradient forces. *See* ATMOSPHERIC GENERAL CIRCULATION.

Because Venus's atmosphere is massive, atmospheric motions are very effective in reducing the horizontal variations of temperature in its lower atmosphere. In the region of the clouds and at lower altitudes, temperature variations occur mostly in the north-south direction, with the equator being only a few degrees warmer than the poles close to the surface and some tens of degrees warmer within the cloud region. Above the clouds, the atmosphere is actually somewhat warmer at the poles than near the equator because of the effects of heat transported by atmospheric motions. As Venus's axis of rotation is almost exactly perpendicular to its orbital plane, its climate has little seasonal variability.

Interior. The similar mean densities of Venus and the Earth imply that Venus is made of rocks similar to those that make up the Earth. However, because Venus formed closer to the Sun, in perhaps a somewhat warmer environment, it may have initially contained a smaller amount of sulfur and water-bearing compounds. Such an environmental difference would probably not have significantly affected Venus's content of the long-lived radioactive elements uranium, potassium, and thorium. Over the lifetime of the Earth, and presumably Venus, the decay of these elements may have generated enough heat to cause these planets to become chemically

Fig. 4. ***Magellan*** **image of a portion of Leda Planitia. The smooth (radar-dark) plains were formed by volcanic lava flows that covered the region. These flows embayed the older, radar-bright, highly fractured or chaotic highlands, sometimes called tessera, that rise out of the plains in the upper-left or northwest part of the image, as well as the circular ring structure in the lower left, which is probably an impact crater. Subsequent volcanism produced radar-bright flows in the upper right. (*Jet Propulsion Laboratory; NASA*)**

differentiated, as free iron melted and sank toward their centers. Also, both planets probably formed "hot" because of gravitational energy released in bringing chunks of rock and planetesimals together to form them. In this case, they may have undergone substantial differentiation in their early histories, with an accompanying release of much of their atmospheric gases. Thus, Venus's interior may be qualitatively similar to that of the Earth in having a central iron core, a middle mantle made of rocks rich in silicon, oxygen, iron, and magnesium, and a thin outer crust containing rocks enriched in silicon in comparison with the rocks of the mantle. However, in contrast to the situation for the Earth, Venus's core may now be either entirely solid or entirely liquid, which could account for the absence of a detectable magnetic field. *See* EARTH INTERIOR.

Rotation. In contrast to the Earth and almost all other planets, Venus rotates in the opposite direction with respect to its orbital motion about the Sun. It rotates so slowly that there are only two sunrises and sunsets per Venus year. Tides raised in the body of the planet by the Sun may have greatly reduced the rate of rotation from an intially large value, similar to the Earth's, to its present low value. Because Venus's atmosphere is so massive, tides raised by the Sun in the atmosphere may have also been important, with its current rate of rotation being determined chiefly by a balance between the oppositely directed torques arising from these two types of solar tides. Venus currently presents the same face to the Earth at the times of closest approach. This suggests that the tidal forces that were exerted by the Earth also played a role and helped lock Venus in its present rotational state.

Life. The current high surface temperature and acid constitution of the clouds preclude the existence of living organisms like those that inhabit the Earth. But Venus may have had a hot ocean of water in its early history when the Sun put out less energy than it does now. If so, it is unclear whether life could have arisen then. Any early life would have been des- troyed when Venus lost its oceans and achieved its current high surface temperature.

James B. Pollack; R. Stephen Saunders

Spacecraft. Venus has been more intensely explored by spacecraft than any other planet. Some 31 (9 of them unsuccessful) United States and Soviet spacecraft have been sent to Venus. The Soviet Union *Venera* and *Vega* series of spacecraft and the United States *Mariner* and *Pioneer* series provided data on the composition of the surface and atmosphere, and the dynamics of the upper atmosphere. *Venera* landers transmitted back the first images from the surface of another planet (**Fig. 2**). The surface of Venus was seen in different places to be rocky with little granular material. The United States *Magellan* by the end of 1992 had mapped nearly 98% of the surface. Venus is the last of the inner planets to be mapped in detail but is the most important planet for understanding how the Earth evolved as a life-sustaining habitat. *See* SPACE PROBE.

Surface. *Magellan*'s mapping mission revealed a unique global volcanic and tectonic style on Venus. Broad volcanic plains make up about 85% of the surface of Venus. The rest is tectonically deformed, higher-standing terrains with complex systems of folds and faults. Regional tectonism is evident in the widespread compressional and extensional deformation of much of the surface material (**Fig. 3**). Venus apparently has had a dynamic mantle that has driven crustal warping, which may be ongoing. However, while various regions of the planet show evidence of motion, no evidence of Earth-like plate tectonics has been found. *See* PLATE TECTONICS.

Long, narrow troughs are seen in many areas where the crust has ruptured; these linear rift zones are associated with extensive broad, domical rises and shield-volcano complexes. Examples are the Sif-Gula region, Beta Regio, and Atla Regio. Large areas of the planet are covered by lava that flowed from volcanic vents (**Fig. 4**). Volcanism on Venus occurs globally, unlike Earth where volcanism is mostly restricted to linear zones that define plate boundaries.

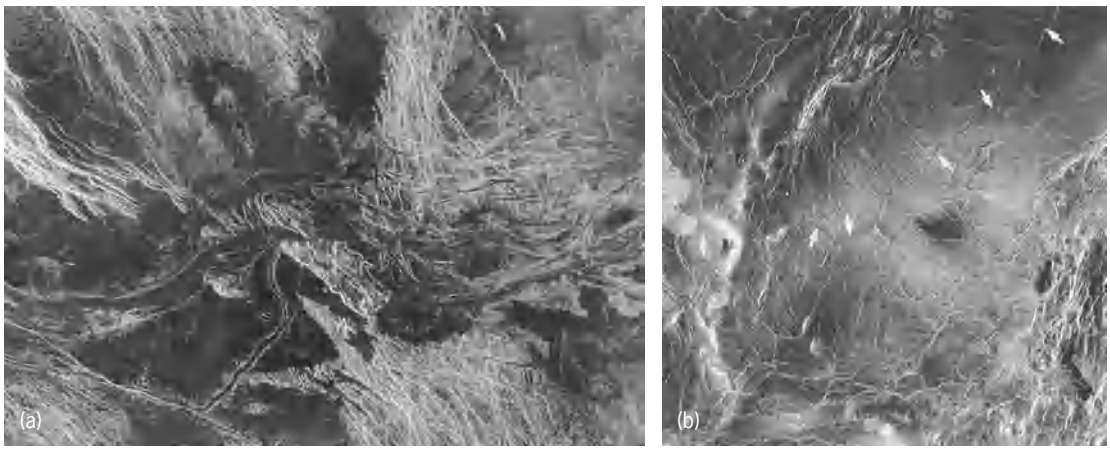


Fig. 5. ***Magellan*** **images of channels on Venus. (*a*) Part of a channel in the Lada Terra region, 750 mi (1200 km) long and 12 mi (20 km) wide, displaying numerous streamlined structures that attest to the very fluid lavas responsible for carving the channel. (*b*) A 360-mi (600-km) segment of the Hildr Channel (indicated by arrows), whose length of 4200 mi (6800 km) makes it the longest known channel in the solar system. (*Jet Propulsion Laboratory; NASA*)**

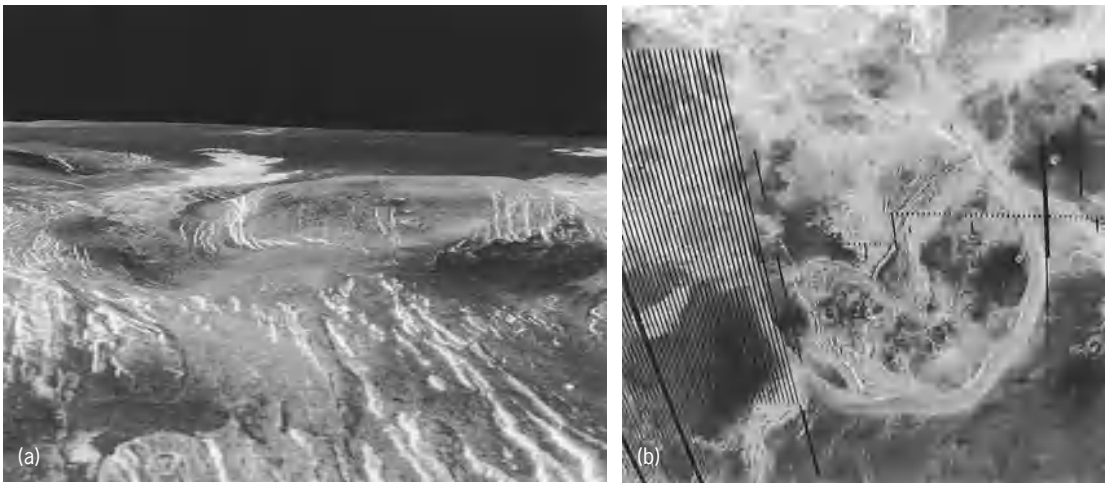


Fig. 6. ***Magellan*** **images of coronae on Venus. (*a*) Computer-simulated view of the corona Idem-Kuva, 60 mi (97 km) in diameter. Lava flows extend for hundreds of miles across the fractured plains in the background. (*b*) Artemis Chasma, whose diameter of 1300 mi (2100 km) makes it the largest corona identified. The interior contains complex systems of fractures, numerous flows, and small volcanoes, and at least two impact craters. The margin forms a steep trough with raised rims. (*Jet Propulsion Laboratory; NASA*)**

The planet also has some unexplained surface features, including long channels meandering across the plains (**Fig. 5***a*). One channel is 4200 mi (6800 km) long (Fig. 5*b*), which is longer than the Nile river. Its origin remains unknown. Many of the channels are clearly formed by lava, but even under the high temperature (850°F or 730 K) of the surface, most known volcanic lava compositions should solidify before they could flow far in open channels. A more exotic lava such as sulfur or a carbonate may have formed some of the longer channels. Another unique landform is the scat- tered flat, circular volcanoes that resemble giant pancakes. These range in diameter from about 9 to 50 mi (15 to 80 km) and are a few hundred meters thick.

Coronae are classified as volcanic-tectonic structures (**Fig. 6**). These features are oval to circular and range in diameter from 60 to 1300 mi (100 to 2100 km). Most are around 190 mi (300 km) in diameter. They have low relief, up to 1–2 mi (2–3 km) high. The margins are generally irregular concentric ridges and troughs. The interiors may be quite irregular, including various tectonic and volcanic edifices. Many coronae have associated volcanic flows. A mantle plume model has been postulated for the formation of coronae, in which the rising lower-density plume causes doming, with contemporaneous or subsequent relaxation to form the bounding ridges and troughs.

The higher elevations of Venus, Maxwell Mons for example, have distinctly higher-than-average radar reflectivity and low emissivity. One explanation for this behavior is that, at slightly lower temperature and pressure in the more elevated regions, a metallic mineral is stable and distributed throughout the surface material to create higher reflectivity. Possible minerals could be pyrrhotite, pyrite, or magnetite.

Impact craters. Impact craters are much less abundant on Venus than on the Moon or Mars. Approximately 950 impact craters have been identified in *Magellan* images (**Fig. 7**). Most appear to be unmodified by erosion. Based on the abundance of impact craters, the average age of the surface appears to be about 300 to 500 million years old. This age would imply a low rate of surface modification by erosional, or even by volcanic and tectonic, processes, and suggests that parts of the planet have undergone relatively recent volcanic or tectonic activity. On a global scale, impact craters are randomly distributed, allowing the determination of the mean surface age of less than 500 million years, but the small number of total crates provides little age information at a local scale.

Asteroids and comets that collide with Venus should have typical velocities of about 12 mi/s

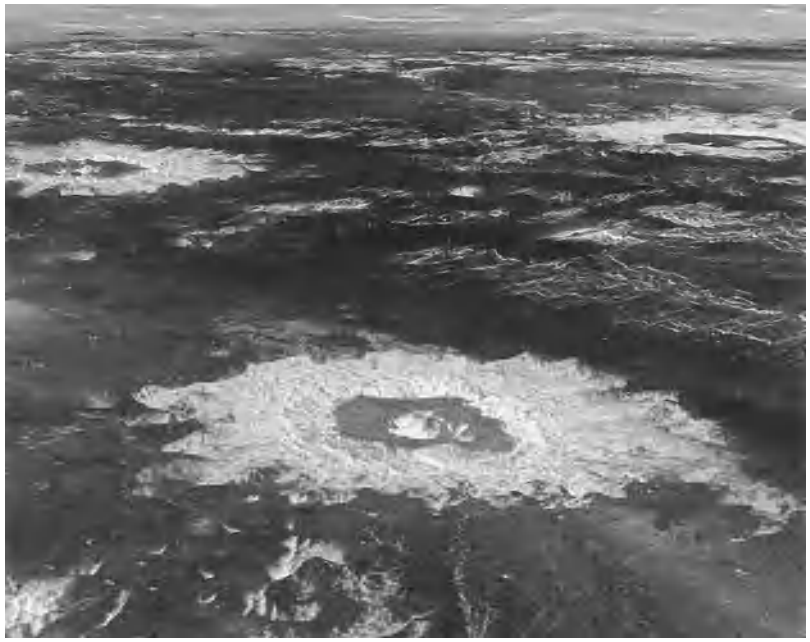


Fig. 7. Computer-simulated perspective view, based on data from ***Magellan*** **spacecraft, showing three impact craters in the northwestern portion of Lavinia Planitia with diameters ranging from 23 to 39 mi (37 to 63 km). (*Jet Propulsion Laboratory; NASA*)**

(20 km/s). Venus's thick atmosphere plays a significant role in meteorite impacts. Because the atmosphere is so dense, only craters larger than 2 mi (3 km) in diameter can form, except in crater clusters where large projectiles apparently broke up before impact, peppering an area with smaller fragments. Ordinarily, a projectile that would produce a crater smaller than about 2 mi (3 km) in diameter would vaporize or break up in its transit to the surface.

Wind features. It was of major scientific interest to discover wind activity on Venus. *Magellan*'s radar saw in many areas abundant bright and dark wind streaks near topographic barriers such as small volcanic ridges. Wind streaks have been found to be most frequent near large impact craters. The craters may have provided the material that is moved by the wind. Also, the impact process itself may have produced some streaks. Some of the larger impact craters have large parabolic features hundreds of miles long. Many of the wind streaks are in the vicinity of these parabolas. The parabolas may be bright or dark, and all are open to the west. This orientation is consistent with some interaction with the east-to-west winds tens of miles above the surface of Venus, which reach speeds of 125 mi/h (200 km/h). Particulate crater ejecta may have been distributed downwind. Alternatively, the turbulent atmospheric disturbance created by the impact may have been carried to the west by the general atmospheric circulation, creating deposition or erosion patterns in the lee of the event. Further study of the orientations of some wind streaks may provide a better understanding of Venus's global wind patterns.

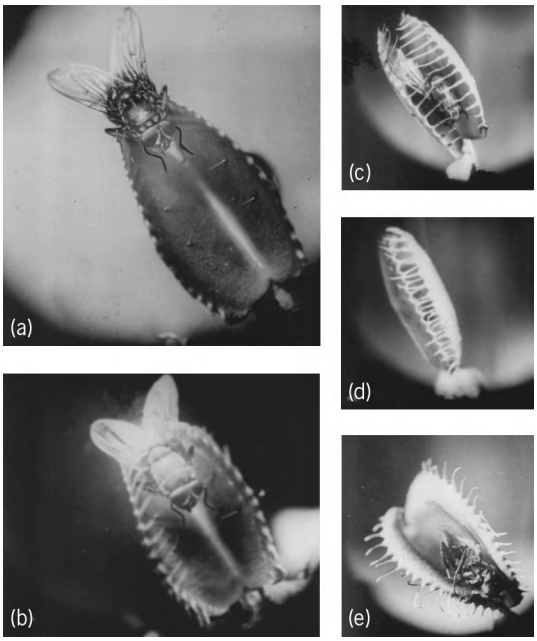
History of volcanism. Currently active volcanism has not been detected on Venus. One theory postulates cessation of a violent episode of global volcanism 300 to 500 million years ago. Subsequently there would have been less volcanic activity, forming the large shield volcanoes such as Maat Mons that are scattered over the planet. It is estimated that the activity represented by the younger volcanoes is less than 5% of the older plains flows. An alternative hypothesis holds that there is an equilibrium between the formation of impact craters and the volcanic outpourings that erase them, so that the surface always appears to be about 500 million years old. *See* PLANET.

R. Stephen Saunders

Bibliography. J. K. Beatty, C. C. Petersen, and A. Chaikin (eds.), *The New Solar System*, 4th ed., 1999; S. W. Bougher, D. M. Hunten, and R. J. Phillips (eds.), *Venus II: Geology, Geophysics, Atmosphere, and Solar Wind Environment*, University of Arizona Press, 1998; P. Cattermole, *Venus: A New Geology*, 1994; H. S. Cooper, Jr., *The Evening Star: Venus Observed*, 1994; D. J. Eicher, *Magellan* scores at Venus, *Astronomy*, 19(1):34–42, January 1991; D. H. Grinspoon, *Venus Revealed*, 1997; *Magellan* reports, *J. Geophys. Res.*, 92:13,059-13,690, Aug. 25, 1992, and 92:15,921-16,382, Oct. 25, 1992; *Magellan* reports, *Science*, 252:247–312, Apr. 12, 1991; *Pioneer Venus* results, *Science*, 203:743–808, Feb. 23, 1979, and 205:41–121, July 6, 1979.

Venus' flytrap

Dionaea muscipula, an insectivorous plant of North and South Carolina (see **illus.**). The two halves of a leaf blade can move as if they were hinged along the midrib and, swinging upward and inward, the two surfaces come together. Any insect alighting on a leaf triggers this sensitive motor mechanism, and is caught between the closing halves of the leaf blade. In this trap, the insect is slowly digested by enzymes secreted by cells in the leaf. *See* INSECTIVOROUS PLANTS; NEPENTHALES; SECRETORY STRUCTURES (PLANT).



Stages *a–e* in capture and digestion of fly by leaf of Venus' flytrap. (*General Biological Supply House*)

Perry D. Strausbaugh; Earl L. Core

Vermiculite

A group of minerals common in some soils and clays and belonging to the family of minerals called layer silicates. Species within the vermiculite group are denoted as either dioctahedral vermiculite or trioctahedral vermiculite, with two or three octahedral cation sites occupied per formula unit, respectively.

Structure. Trioctahedral vermiculite is a 2:1 layer silicate with a fundamental unit similar to that of mica. An octahedral sheet forms the basis of the layer. This sheet is composed of octahedra (a polygon with six corners and eight faces) with magnesium (Mg) or similar-sized cation [iron (Fe), aluminum (Al)] as the central cation and six surrounding oxygens (O) and hydroxyls (OH). To form the 2:1 layer, the octahedral sheet is sandwiched between two opposing tetrahedral sheets. Tetrahedra consist of T cations [T = silicon (Si) or aluminum] surrounded by four oxygen atoms. *See* MICA; PHLOGOPITE.

Unlike micas, however, vermiculites have an additional octahedral sheet between the 2:1 layers. This sheet is partially occupied by magnesium [or lesser amounts of sodium (Na) or calcium (Ca)] and water (H_2O) molecules, but both occupy definite sites. The sheet has not been determined precisely. The stacking of layers in vermiculite, although potentially very complex, always occurs as a two-layer form; semirandom stacking sequences are the norm. The structure resembles chlorite also, but the interlayer octahedral sheet in vermiculite is not as fully occupied as in chlorites. *See* CHLORITE; HYDROGEN BOND.

Properties. Perfect basal cleavage develops by the layerlike structure. Density varies but is near 2.4 g/cm^3; hardness on the Mohs scale is near 1.5; and luster is pearly. Thin sheets deform easily and may be yellow to brown. Rapid heating causes some vermiculites to undergo a "popcorn" effect, where interlayer water explosively discharges to cause exfoliation, whereas other vermiculites exfoliate more slowly into wormlike threads. *See* HARDNESS SCALES.

Occurrence. Trioctahedral vermiculites form as an alteration product of phlogopite, biotite, or chlorite, whereas dioctahedral vermiculites are generally thought to form from illite. Trioctahedral vermiculites are common in well-drained soils in humid regions wherever the parent material contains trioctahedral micas. Interstratification of vermiculite and mica is common. *See* BIOTITE; ILLITE; MUSCOVITE.

Uses. Heat-treated and expanded vermiculite is used as an insulator in construction. Mixed with plaster and cement, vermiculite is used to make lightweight versions of these materials. Vermiculite is also useful as an absorbent for some environmentally hazardous liquids. *See* CLAY, COMMERCIAL; CLAY MINERALS; SILICATE MINERALS. Stephen Guggenheim

Bibliography. S. W. Bailey (ed.), *Hydrous Phyllosilicates* (*Exclusive of the Micas*), 1988.

Vernalization

The induction in plants of the competence or ripeness to flower by the influence of cold, that is, temperatures below the optimal temperature for growth. Vernalization thus concerns the first of the three phases of flower formation in plants. In the second stage, for which a certain photoperiod frequently is required, flowers are initiated. In the third stage flowers are unfolded.

Phenomena of vernalization were reported sporadically in the nineteenth century; however, intensive investigations in this field started with G. Gassner in 1910. The cold requirement for the induction of ripeness to flower is known for a large number of perennial, winter annual, and late-flowering summer annual plants. Examples are winter cereals and other grasses (strains of *Lolium, Bromus, Alopecurus, Phleum, Cynosurus*, and *Dactylis*), beets and cabbages from the genera *Beta* and *Brassica*, and species of *Lupinus, Vicia, Oenothera, Streptocarpus, Daucus, Carum, Apium, Petroselinum, Echium, Campanula, Cichorium, Scorzonera*, and *Helianthus*. These plants are prevented from floral initiation before winter by the cold requirement. Most physiological work on vernalization has been done with winter rye, henbane (*Hyoscyamus niger*), *Chrysanthemum morifolium*, and *Arabidopsis thaliana*.

Normally, the shoot apex of either a plant or an embryo must be exposed to cold for vernalization to occur. Root cuttings of *Lunaria biennis* and of *Cichorium intybus* and leaf cuttings of *Lunaria*, which produce shoot meristems in their adventitious roots, may be vernalized successfully, too, in certain cases even before the formation of shoot meristems. In unifoliate *Streptocarpus* species, vernalization takes place in the macrocotyledon.

Many plants can be vernalized in the seed stage. However, some biennial and perennial plants like *Hyoscyamus niger, Dianthus barbatus*, and some strains of *Brassica oleracea* do not respond to cold treatment during a juvenile phase, which is usually 1–3 months. In certain races of winter wheat, lettuce, and *Arabidopsis* the vernalizibility diminishes during germination, or the first days or weeks thereafter, and then increases again (that is, the vernalizing effect of a given cold treatment is relatively great in seeds, smaller in young seedlings, and greater again in older plants). Vernalization of seeds is only possible after imbibition; between 30 g of water for caryopses of some cereals and 100 g for certain papilionaceous seeds must be added to 100 g of seeds. Caryopses of cereals may even be vernalized prior to maturity, when still upon the mother plant.

Cold duration and temperature. The minimum duration of a cold treatment that will barely accelerate flower formation is 4 days in Petkus winter rye, 5 days in a Japanese radish variety, and 7 days in biennial *Hyoscyamus niger*. The optimal duration of vernalization varies. It is less than 3 weeks for the radish and those late summer cereals requiring a small amount of cold and more than 100 days for strains of *Hyoscyamus niger, Lactuca sativa*, and *Lupinus albus*.

In some instances, such as winter rye, *Hyoscyamus*, and *Arabidopsis*, the longer a vernalization treatment lasts, the lower is the optimal temperature of vernalization. It then approaches temperatures just above zero. The upper limit of vernalizing temperatures is known exactly only in some plants that have high cold requirements, such as a variety of *Allium cepa* (59–68°F or 15–20°C), sugarbeet (50–52°F or 10–11°C), *Hyoscyamus* (63°F or 17°C), and kohlrabi (about 50°F or 10°C). Temperatures below zero (down to 18°F or −8°C) have also proved effective in some cases, such as caryopses of winter cereals and *Arabidopsis* seeds, but to a rather small degree.

Most cold-requiring plants are also frost-resistant; their frost resistance, however, usually diminishes during vernalization for reasons which are not completely known. *See* COLD HARDINESS (PLANT).

Effect of photoperiodism. The vernalization effect is measured by time in days to anthesis, to ear emergence (in grasses), and to unfolding of the first flower and by the number of leaves on the main shoot axis

below the insertion of the first flower. The effect is influenced by photoperiod, light intensity, or both before, during, and after vernalization. The vernalization effect is increased when plants are exposed prior to cold treatment to low light intensity (*Arabidopsis thaliana*), to short-day conditions (*Daucus carota* and grasses like *Poa pratensis, Bromus inermis*, and *Dactylis glomerata*), or to long-day conditions (short-day plant *Chrysanthemum morifolium* var. Sunbeam). Illumination during vernalization has opposite effects in different species. In many long-day plants a vernalization effect is manifested only when vernalization is followed by long-day exposure (*Daucus carota*, many winter cereals and other grasses, and strains of *Arabidopsis* requiring much cold); in others, the greatest relative vernalization effect occurs under short-day conditions (*Beta vulgaris, Agrostemma githago*, and *Arabidopsis* strains requiring less cold). *See* PHOTOPERIODISM.

Devernalization. The vernalization effect can be annulled (devernalization) immediately after cold treatment by high temperatures (several days at 77–104°F or 25–40°C) if the preceding cold exposure did not last too long (winter rye and *Dianthus barbatus*), if it took place in darkness (*Arabidopsis*), and if the plants were not too old at the beginning of the cold treatment (*Arabidopsis*). Short days or low light intensity may have devernalizing effects, too (sugarbeet, *Oenothera biennis*, and perennial plants like certain chrysanthemums). Devernalized plants can be revernalized.

Antivernalization. Flowering may be delayed and the effect of subsequent vernalization may be reduced by a heat treatment, for example, 3–5 days at 86–104°F (30–40°C) for lettuce, *Arabidopsis*, and winter rye. This antivernalization is applied to onions in order to prevent bolting.

Genetic basis. Up to six genes have been found responsible for the cold requirement of vernalizable plants. In the case of three of these genes—*Hb, Ii*, and *Kk*, which are found in wheat, barley, rye, and *Arabidopsis*—the cold requirement is caused by the recessive alleles; in the case of the other genes—*Ss, Tt*, and *Uu*, found in wheat, barley, *Hyoscyamus*, and *Arabidopsis*—it is caused by the dominant alleles. *I* and *K* (found in barley and wheat) are epistatic to *S*, and *ss* is epistatic to *ii* and *kk*; *T* (found in *Arabidopsis*) is epistatic to *H* and *hh*. A vernalization effect is never transmitted to sexual descendants. *See* GENE ACTION.

Biochemical basis. These differences in the genetic basis of vernalization suggest that vernalization has different biochemical backgrounds in different species. Thus the results gained by application of physiological chemical methods do not form a clear picture. In winter rye vernalization occurs only in the presence of oxygen. In some cold-requiring plants, like *Hyoscyamus* and *Arabidopsis* (but not in many Campanulaceae and Gramineae), cold can be replaced by gibberellins, which may overcome an NH_3-induced inhibition of vernalization in winter wheat; in others, like sugarbeet and *Oenothera*, gibberellins can replace cold only partially. Gibberellins may have certain side effects (for example, on internode elongation and ramification), which vernalization does not have, and they are almost never effective on the seed stage. In winter wheat (but not in winter rye), vernalization can perhaps be specifically inhibited by maleic hydrazide, which promotes flower formation in cold-requiring celery, and by chloramphenicol, which increases the effect of incomplete vernalization in winter rye. In the radish vernalization can be replaced to a certain degree by ribonucleic acid (RNA), and in *Streptocarpus* and *Oenothera* vernalization can be inhibited by thiouracil and fluorouracil, respectively, which replace uracil in RNA synthesis. However, no correlation between RNA, deoxyribonucleic acid (DNA), and vernalization has been found in other cases such as winter rye and *Hyoscyamus*. Studies of the role of auxins in vernalization also have given varying results.

Vernalization may further be partially or entirely replaced in certain plants by vitamin E (winter rye, *Cichorium intybus*), kinetin (*Cichorium intybus* and *Geum urbanum*), decapitation (*Chrysanthemum morifolium* and *Geum urbanum*), heat (*Scrophularia alata*), high light intensity (winter cereals and other winter annual grasses and *Geum urbanum*), and by others. Although the vernalization effect is transmissible by grafting in *Hyoscyamus*, it has not been possible to extract from vernalized plants flower-inducing principles other than gibberellinlike substances which, for several reasons, do not seem to be identical with the end product of vernalization. *See* AUXIN; GIBBERELLIN; NUCLEIC ACID; PLANT GROWTH.

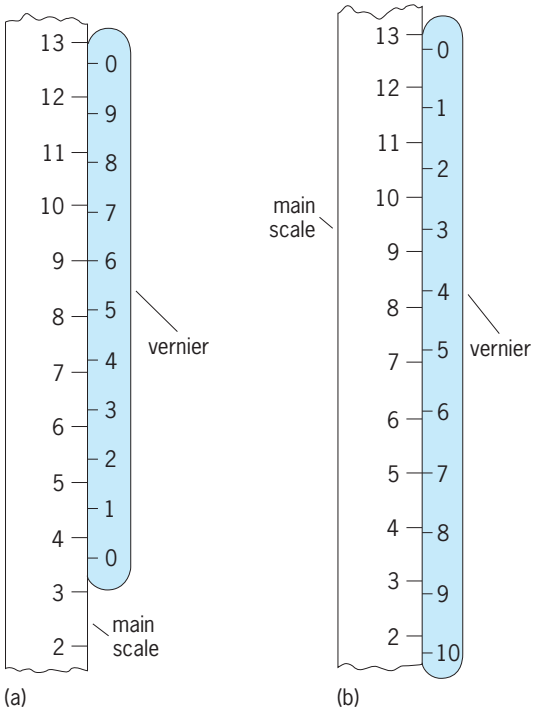
Klaus Napp-Zinn

Bibliography. J. W. Einset, *Plant Growth and Development*, 1988; D. E. Fosket, *Plant Growth and Development: A Molecular Approach*, 1994; P. S. Nobel, *Plant Physiology*, 1991; W. Ruhland (ed.), *Encyclopedia of Plant Physiology*, vols. 15 and 16, 1965; P. F. Wareing and I. D. Phillips, *Growth and Differentiation in Plants*, 3d ed., 1981.

Vernier

A short auxiliary scale placed along the main instrument scale to permit accurate fractional reading of the least main division of the main scale, invented by Pierre Vernier about 1630. The auxiliary, or vernier, scale is graduated in one or both directions from the fiducial (index) mark in numbered divisions which are fractionally shorter (in a direct vernier) or longer (in a retrograde vernier) than those on the main scale. The position of the fiducial mark (the zero mark of the vernier scale) between divisions on the main scale is indicated by the number of the graduation on the vernier scale which lines up exactly with a graduation on the main scale.

Each vernier division of the direct-reading vernier in the **illustration** is equal to 0.9 of the main scale divisions. The fiducial mark is located between 3 and 4 on the main scale, and since graduation 6 lines up with the main scale the exact reading is 3.6 ($3.6 + 6 \times 0.9 = 9.0$).



Two types of vernier scales. (*a*) Direct (reading 3.6). (*b*) Retrograde (reading 12.7).

Such a vernier scale gives a direct reading to the nearest tenth division and permits estimation, if none of the graduations on the vernier line up exactly, to 0.03 or 0.05 divisions.

Vernier scales are used in both linear and circular measurements, as on calipers, screw micrometers, and protractors on telescope mounts and surveying instruments. The divisions are usually tenths or hundredths of units, but may be quarters, twelfths, or other fractions for circular measurements. If the main scale is graduated in half degrees, the vernier divisions may be $^{11}/_{12}$ or $^{29}/_{30}$ of the main scale divisions, to permit reading to 5 or 2 minutes of arc.

The term vernier is often loosely applied to any fine adjustment, for example, to the one of two or more controls, such as valves or rheostats of different range, that causes the least control action per unit motion or other input. William A. Wildhack

Vertebra

The basic unit of the vertebral column. Collectively, the vertebrae surround and protect the spinal cord and provide some type of axial support for the body. The stresses that the vertebral column must meet change somewhat from one end of the animal to the other, and also differ greatly between aquatic and terrestrial vertebrates because the problems of support and locomotion in these media are quite different. Accordingly, vertebral structure varies widely; yet all vertebrae have many features in common.

Basic structure. A vertebra from the thoracic region of a mammal illustrates the basic morphology well (**Fig. 1**). The ventral portion consists of a disc-shaped mass of bone known as the body or centrum. An arch of bone, the neural arch, extends dorsally from the centrum and encompasses a space, the vertebral canal, in which the spinal cord lies. The bases, or roots, of the arch are narrower than other parts so that clefts, the intervertebral foramina, lie between the arch bases of successive vertebrae. Spinal nerves pass through these foramina.

Certain muscles and ligaments attach onto the spinous process, which extends dorsally from the top of the arch, and onto a pair of transverse processes, which extend laterally from the arch. One pair of articular processes, or zygapophyses, extend forward from the neural arch, and another pair extends posteriorly. Articular processes of successive vertebrae overlap and help to hold the vertebrae together. Additional articulating or reinforcing processes are present in certain other terrestrial vertebrates. The centra of adjacent vertebrae are also joined together by invertebral discs of fibrocartilage. Numerous ligaments interlace the vertebrae.

Ribs articulate onto the thoracic vertebrae. Typically, each mammalian rib has two articular surfaces—a terminal head and a tubercle situated a short distance distal to the head. The tubercle articulates with a facet located on the end of the transverse process; the head usually articulates intervertebrally onto the intervertebral disc and adjacent parts of the bordering centra.

Development. The vertebrae develop from mesenchymal cells that migrate from the sclerotomal portions of the somites and spread around the embryonic neural tube and notochord (**Fig. 2**). The first sign of vertebrae differentiation in this mesenchymal sheath is the development of small blocks of cartilage known as the dorsal arcualia. Since the spinal ganglia and nerves have previously begun to differentiate in a segmental position directly adjacent to the muscle segments, the dorsal arcualia, which are the

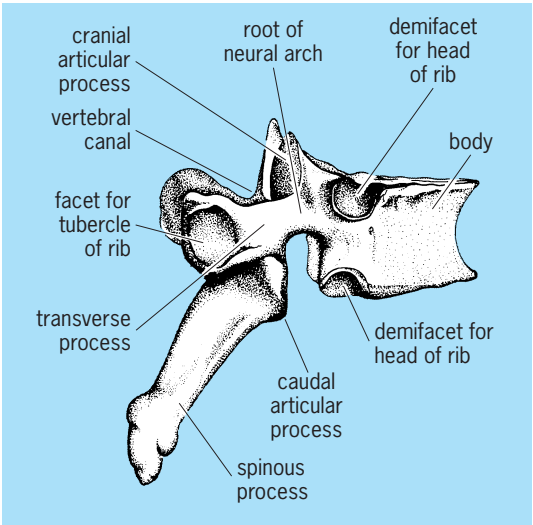


Fig. 1. Lateral view of a human thoracic vertebra. Anterior is toward the right. (*After C. M. Goss, ed., Gray's Anatomy of the Human Body, 28th ed., Lea and Febiger, 1964*)

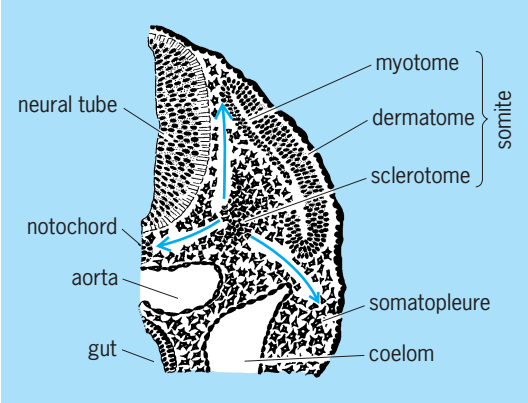


Fig. 2. Diagrammatic cross section through one side of the back of a 4-week-old human embryo. Migration of mesenchymal cells from the sclerotome is shown by arrows. (*After L. B. Arey, Developmental Anatomy, 6th ed., Saunders, 1954*)

rudiments of the neural arches, assume an intersegmental position. The rest of each vertebra, in turn, differentiates intersegmentally, where it can be acted on by muscles of adjacent segments. Most vertebrates have only one pair of dorsal arcualia per body segment, but sharks and other cartilaginous fishes have two pair. A pair of ventral arcualia differentiate adjacent to the ventral part of the notochord and develop either into ribs or, in the tail region of many vertebrates, into a ventral hemal arch that surrounds the caudal artery and vein.

The centra of the vertebrae develop around the notochord, or within its sheath, primarily from mesenchyme derived directly from the sclerotomes. In some fishes, however, bases of the arcualia contribute to the centra. Centra formation is very incomplete in cyclostomes and lungfishes, and a large notochord persists into adult life; but in most vertebrates the notochord is replaced, or nearly so, by the developing centra. A thin strand of notochord runs through the centra in cartilaginous fishes and many salamanders, and traces of the notochord may lie between successive centra. There is a trace within the intervertebral disc of mammals.

The arches and centra remain cartilaginous throughout life in cartilaginous fishes. Parts of the vertebrae may remain cartilaginous in other groups, but by the adult stage of most vertebrates the cartilage is replaced by bone.

Fishes. Since shortening of the body of fishes is resisted by the incompressible, yet flexible, vertebral column, contraction of the longitudinal fibers in the muscle segments, first on one side of the body and then on the other, causes the body to bend from side to side. The vertebral column of fishes is an important part of the locomotor system. Its role in physical support of the body is rather negligible because flesh is only slightly denser than water and needs little support in an aqueous medium. The centra of fishes are strong discs or biconcave cylinders with traces of the notochord persisting in the concavity between them. They resist compression forces well, yet permit the body to bend from side to side. Neural arches lack articular process because there are few vertical forces to be resisted.

Since the entire vertebral column of fishes is performing essentially the same function, there is little regional differentiation. Trunk vertebrae bearing ribs extend from the trunk to the tail. The first vertebra is slightly modified for articulating with the skull, but head and trunk do not move independently and there is no neck. Ribs disappear in the tail region, and the ventral arcualia form an arch of bone, the hemal arch, that encases the caudal artery and vein.

Amphibians. Terrestrial vertebrates receive little support from the surrounding air, and most propel themselves by thrusts of the limbs against the ground. The vertebral column becomes essentially a beam that suspends the body weight and transfers this to the ground by way of the limbs. The vertebral column must resist considerable vertical forces. The centra are thoroughly ossified, and the neural arches are securely united by articular processes.

There is only a little more regional differentiation in the vertebral column in ancestral amphibians than in fishes. A single vertebra is specialized for articulating with the head; no well-defined neck region is present. A series of dorsal, or trunk, vertebrae extend to the pelvic girdle. Here a single sacral vertebra and pair of ribs are specialized for articulating with the pelvic girdle. Anteriorly in terrestrial vertebrates weight is transferred from the vertebral column and ribs to the pectoral girdle by a muscular sling. Caudal vertebrae follow the sacral vertebra. Ribs articulate with all vertebrae from the head into the base of the tail. Longest near the front of the series, they are progressively shorter posteriorly.

Mammals. Regional differentiation of the vertebral column continues in reptiles, and by the mammalian stage of evolution different parts of the column are clearly specialized to subserve certain functions in addition to their general supportive role.

Cervical vertebrae. The head moves independently of the trunk, and a distinct neck region is present. With few exceptions all mammals from a shrew to a giraffe have seven cervical vertebrae. In whales and porpoises, which have readapted to an aquatic mode of life by assuming a fishlike body form, an independent motion of head and trunk would be undesirable. Seven cervical vertebrae are present, but they are much shortened and fused together. Small embryonic ribs have fused on the sides of the cervical vertebrae and contribute to the transverse process. Another diagnostic feature of most cervical vertebrae is the perforation of the transverse process by a transverse foramen that carries the vertebral artery and vein (**Fig. 3**). The first two cervical vertebrae, the atlas and axis respectively, are so modified as to permit the head to move in nearly any direction.

Thoracic vertebrae. Well-developed ribs, which play an important role in respiratory movements, articulate with the anterior trunk or thoracic vertebrae of mammals. The more anterior ribs are quite short and stout and also play an important role in weight transfer from the vertebral column to the pectoral girdle. A strong muscular sling, composed primarily of the

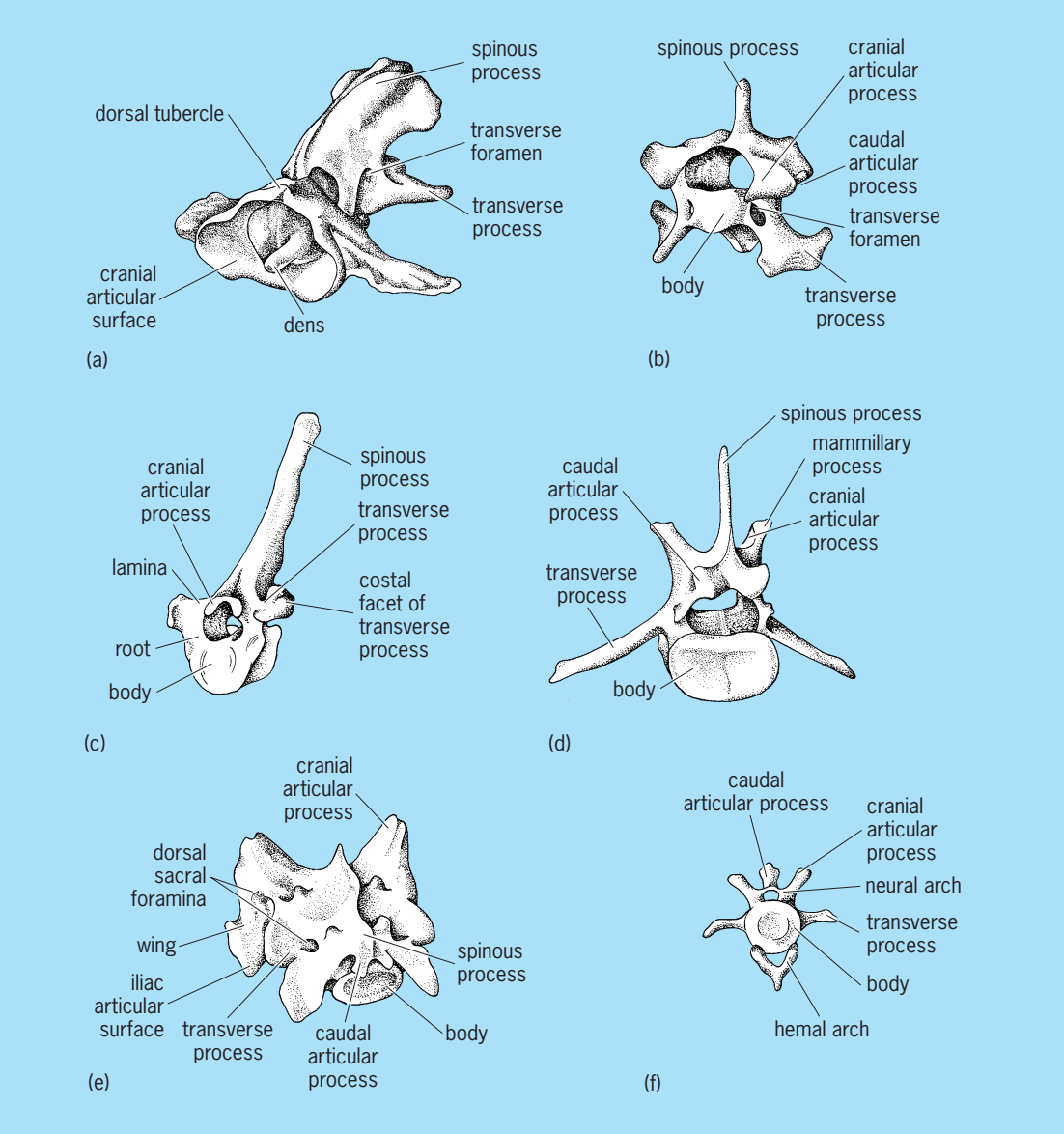


Fig. 3. Representative vertebrae of a dog. (*a*) Atlas and axis articulated, cranial lateral aspect. (*b*) Fifth cervical, cranial lateral aspect. (*c*) Sixth thoracic, cranial lateral aspect. (*d*) Fifth lumbar, caudal lateral aspect. (*e*) Sacrum caudal, lateral aspect. (*f*) Fourth caudal, cranial aspect. (*After M. E. Miller, G. C. Christensen, and H. E. Evans, Anatomy of the Dog, Saunders, 1964*)

serratus ventralis, extends from the ribs to the dorsal border of the scapula. Mammalian thoracic vertebrae, and only the thoracic vertebrae, bear distinct articular surfaces for ribs. The number of thoracic vertebrae varies between species but is on the order of 11 (bat) to 18 or 20 (horse). Humans have 12.

Lumbar vertebrae. Lumbar vertebrae occupy the posterior part of the mammal trunk region. They are characterized by relatively large transverse processes to which certain powerful back muscles attach. The transverse processes include embryonic rib rudiments that have fused onto the vertebrae. The number of lumbar vertebrae vary between species but is on the order of 5 (bat) to 8 (whale). Humans have 5.

Sacral vertebrae. Correlated with the greater efficiency of terrestrial locomotion and the need for strong support for the powerful hindlegs, the number of sacral vertebrae increased during evolution from the single one of amphibians. Reptiles usually have two, and most mammals have three which are fused together, along with their embryonic rib rudiments, into a solid complex of bone called the sacrum. In species in which there are heavy stresses in this region, the number of sacral vertebrae is increased. Humans, which are bipeds, have 5, and certain of the powerful hoofed mammals, for example, the horse, also have 5. Birds, whose hindlegs act as shock absorbers upon landing, have between 10 and 23 vertebrae fused together in their synsacrum.

Caudal vertebrae. The tail and caudal musculature no longer play an important role in the locomotion of most mammals (the Cetacea being a conspicuous exception), and the tail is greatly reduced in size. The spinal cord of mammals ends within the lumbar region, and only a few spinal nerves continue through the vertebrae canal into the tail. Caudal vertebrae are

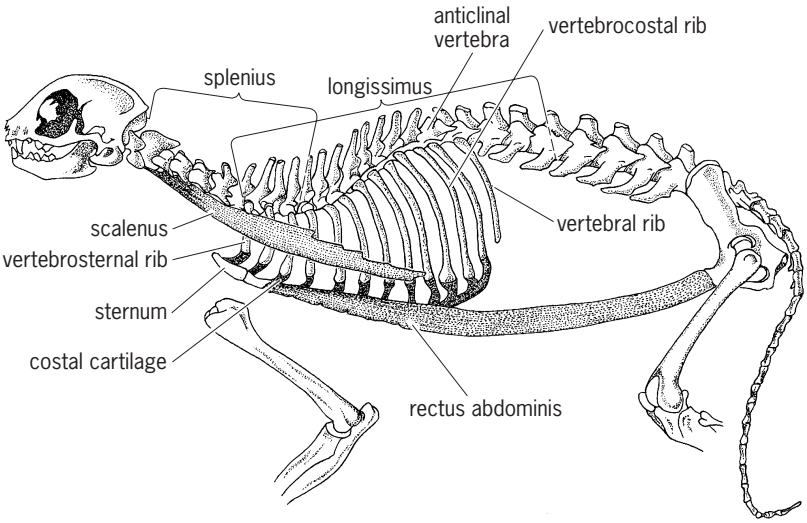


Fig. 4. Lateral view of cat skeleton to illustrate Slijper's views on biomechanics of trunk skeleton. The brackets indicate lines of action of certain muscles upon the neural spines. (*After W. F. Walker, Vertebrate Dissection, 4th ed., Saunders, 1970*)

small and become progressively incomplete as one moves distally along the tail until only centra are left. The more proximal caudal vertebrae often bear on their ventral surfaces small, V-shaped hemal arches or chevron bones. Tail length, and hence the number of caudal vertebrae, vary widely. Some oppossums have as many as 35, and humans, in which the tail is absent as an external structure, have only 3 to 5 caudal vertebrae. These form an internal coccyx to which certain anal muscles attach.

Biomechanics. The vertebral column of mammals is a complex structure that forms the primary girder of the skeleton, subserves the functions discussed above, and by its extensions and flection takes part in the locomotion of certain fast-moving quadrupeds. In order to analyze its biomechanics, some authors have compared it to a cantilever girder, but others have taken a broader view and have compared the entire trunk skeleton and its muscles to an archer's bow (**Fig. 4**). The centra form the main supporting "bow." It is under tension from dorsal muscles and would tend to straighten out if it were not for the "bowstring" (pelvic girdle, rectus abdominis, and certain ribs) counteracting this force. This hypothesis views the vertebral column as a dynamic girder capable both of providing support and of participating in the movements of the body.

The spinous processes are interpreted as muscle lever arms acting on the centra. The direction of their inclination tends to be perpendicular to the major muscle forces acting upon them. The increase in height of the spinous processes at the base of the neck and front of the thorax increases the length of these lever arms and the mechanical advantage of the muscles acting on them. Many of these muscles act to support the relatively heavy head on the end of the vertebral beam. *See* BIOMECHANICS; SKELETAL SYSTEM.

Warren F. Walker, Jr.

Bibliography. R. L. Carroll, *Vertebrate Paleontology and Evolution*, 1987; A. S. Romer and S. T. Parsons, *The Vertebrate Body*, 6th ed., 1986.

Vertebrata

The largest natural group (clade, monophyletic group) of chordate animals. Like all chordates, vertebrates have a notochord, but it is largely replaced by vertebrae (initially neural arches that surround the nerve chord) during development. Together with the living hagfishes (Hyperotreti; Myxiniiformes) and many fossil groups, vertebrates are members of the Craniata. One major theme of craniate evolution is the increasing elaboration of the brain, its cranial nerves, and many sensory organs, such as the organs of smell (olfactory), hearing (inner ear) and feel (sensory organs of the lateral line system). A unique innovation is the evolution of neural crest cells, cells that begin development along the neural crest of the embryo and which then migrate to many parts of the body, giving rise to or contributing to the formation of such structures as the cranium and the branchial arches (a remnant of which remains in humans as parts of the larynx). Vertebrates have a long fossil history. Earliest fossils of jawless vertebrates are known from the late Cambrian (490 million years before present, mybp). Jawed vertebrate fossils are known from the Ordovician (450 mybp) and this group began to predominate in the Devonian (about 360 mybp). *See* CHORDATA; NERVOUS SYSTEM (VERTEBRATE).

Formal vertebrate classifications of the past emphasized overall distinctiveness and grade of organization. Modern (evolutionary phylogenetic) classifications are based on common ancestry and genealogy. Thus, traditional classifications have been highly modified and classifications in many textbooks are out of date. Older classifications recognized a group, Agnatha, which contained lampreys and the more primitive hagfishes. Modern classifications recognize the evidence that lampreys, although jawless, are nevertheless more closely related to jawed vertebrates than to hagfishes. Older classifications placed the lungfishes and coelacanth with the bony fishes. Modern classifications place them with the legged vertebrates, because they share a common ancestor with legged vertebrates not shared with bony fishes. *See* ANIMAL SYSTEMATICS; SARCOPTERYGII.

Lampreys and fossil jawless fishes. The most basal living members of Vertebrata are the lampreys (Hyperoatia; about 35 species). Lampreys are jawless fishes whose eel-like body resembles hagfishes, but whose predatory, round, and toothed mouth is unique among vertebrates. Lampreys and many fossil jawless vertebrates that flourished in the Paleozoic have a number of unique evolutionary innovations. These include neural arches (vertebrae) that enclose the dorsal nerve chord, muscles that work to move the fin rays, vertebrate heart with atrium and ventricle closely situated, nervous regulation of the heart, and a number of important sensory innovations including the evolution of extrinsic eye muscles, the elaboration of the inner ear to two vertical semicircular canals (later to be three in jawed vertebrates), and the development of true neuromasts (sensory

hair organs) in the sensory canal system. Of course, some of the soft anatomy cannot be observed in fossils.

Jawed vertebrates. Major evolutionary innovations characterized the jawed vertebrates, Gnathostomata ("jawed mouth"). In addition to jaws and characters mentioned above, gnathostomes have paired fins or the homologue of paired fins, legs. They also have a mineralized exoskeleton that surrounds certain cartilages of the head and fins and consists of true bone (as in human leg bones) or calcified cartilage (as in many sharks). As is usual in evolution, such characters evolve stepwise; there are jawless relatives of gnathostomes with paired fins, and more primitive relatives that have mineralized bone but which lack either paired fins or jaws. One of the remarkable innovations of the gnathostome ancestor was the evolution of the adaptive immune system, the basis of our ability to fight off infection and to develop vaccines to fight many infectious diseases.

Placodermi and Chondrichthyes. There are two major living clades of gnathostomes and one major extinct clade. The extinct clade is Placodermi, a group of Paleozoic fishes characterized by heavy body armor and which includes some of the largest predatory fishes of Paleozoic waters. The smaller of the living clades is Chondrichthyes. It includes the chimeras (30 species) and sharks and rays (about 830 species). Chondrichthyes has a rich fossil history; the earliest remains are known from the Late Ordovician, and early sharks were common predators in both Paleozoic and Mesozoic seas. They are characterized by several evolutionary innovations including internal fertilization via the unique male clasper organs, and the histology of the calcified cartilage that surrounds the cartilaginous skeleton (micrometric prismatic cartilage). The lack of an obvious bony skeleton is a secondary innovation of sharks and is not primitive. In fact, Chondrichthyes fishes have bone, which form the base of each scale. *See* CHONDRICHTHYES; PLACODERMI.

Osteichthyes. The third major clade of vertebrates is Osteichthyes, a group that in traditional classification contained only the bony fishes, but which now includes the Actinopterygii (bony fishes, about 25,000 species) and the Sarcopterygii (about 24,000 species). Sarcopterygii includes the legged vertebrates (amphibians, reptiles, mammals, and birds), the living lungfishes, coelacanths, and their diverse fossil relatives commonly known as crossopterygians. *See* OSTEICHTHYES.

Current classification. A current classification of living vertebrates is shown below. Note that the ranks of this classification serve only to organize and subordinate the groups; ranks do not imply distinctiveness or uniqueness of adaptations. Many of the traditional classes, such as class Mammalia, would have lower rank in modern classifications such as this one.

Subphylum Chordata
  Infraphylum Craniata
    Superclass Hyperotreti (hagfishes)
      Vertebrata
        Class Hyperoatia (lampreys)
          Gnathostomata
            Infraclass Chondrichthyes (chimeras, sharks, rays)
            Osteichthyes
              Subclass Actinopterygii (bony fishes, sturgeons, trouts, basses, tunas, etc.)
              Sarcopterygii (lungfishes, coelacanths, amphibians, reptiles, birds, mammals)

E. O. Wiley

Vertebrate brain (evolution)

A highly complex organ consisting of sensory and motor systems that constitutes part of the nervous system. Virtually all of the brain systems that are found in mammals occur in birds, reptiles, and amphibians, as well as in fishes and sharks. These systems have become more complex and sophisticated as the adaptive requirements of the animals changed. Occasionally, new sensory systems arose, most likely as specializations of existing systems; however, some of these disappeared as animals left the aquatic world. Some sensory systems arose and declined several times in different lineages. In spite of the many changes in brain structure, some of them quite dramatic, the evolution of the nervous system, from the earliest vertebrates through to those of today, has been relatively conservative. *See* NERVOUS SYSTEM (INVERTEBRATE); NERVOUS SYSTEM (VERTEBRATE).

Central nervous system. The nervous system consists of two main divisions: the central nervous system, which is made up of the brain and the spinal cord, and the peripheral nervous system. The peripheral nervous system consists of the nerves that bring information from the outside world via the eyes, ears, nose, skin, and other sensory systems, and the nerves that carry information from the body's interior to the spinal cord and brain. These nerves also convey commands from the brain and spinal cord to the external muscles that move the skeleton, as well as to various internal organs and glands. *See* CENTRAL NERVOUS SYSTEM.

The head end of any bilaterally symmetrical animal is the most important end. Most of the sensory systems are located in or on the head because that is the ideal placement for systems that monitor the environment for danger, something good to eat, a potential mate, a safe place to hide, and anything else important to the animal's survival.

The brain consists of a variety of systems, some of which are sensory and deal with the acquisition of information from the internal and external environments. Other systems are motor and are involved with the movement of the skeletal muscles; the muscles of the internal organs, such as the heart and the digestive and respiratory systems; and the secretions of certain glands, such as the salivary and tear glands. The bulk of the brain, however, is composed of systems that are neither sensory nor motor, although

they have strong connections with both. Instead, these systems are integrative and organize, coordinate, and direct the activities of the sensory and motor systems. These integrative systems regulate such processes as sleep and wakefulness, attention, the coordination of various muscle groups, emotion, social behavior, learning, memory, thinking, planning, and other aspects of mental life. Social behavior itself is highly complex and includes such interactions between individuals as courtship, mating, parental care, and the organization and structure of groups of individuals. *See* BRAIN; MOTOR SYSTEMS.

Development. The nervous system develops in the embryo as a hollow tube. The remnants of this hollow tube in the adult are known as the brain ventricles and are filled with the cerebrospinal fluid. Two types of overall brain organization are found among vertebrates based on the relationship of the neurons to these ventricles. In brains with the laminar type of organization, the neurons have not migrated very far from the layer immediately surrounding the hollow ventricular core of the brain, which is the zone in which neurons are born (**Fig. 1**). This type of organization is typical of amphibian brains and is also found in the brains of some sharks, fishes (especially those in which their skeletons are partly or mostly made of cartilage rather than bone), and lampreys (jawless fishes). In contrast, in the brains of those vertebrates with the elaborated type of organization, the neurons have migrated from the zone around the ventricles to occupy nearly all of the interior of the brain. This type of brain organization is typical of reptiles, mammals, birds, fishes (with fully bony skeletons), as well as skates, rays, some sharks, and hagfishes (jawless fishes) [**Fig. 2**]. *See* NEURON.

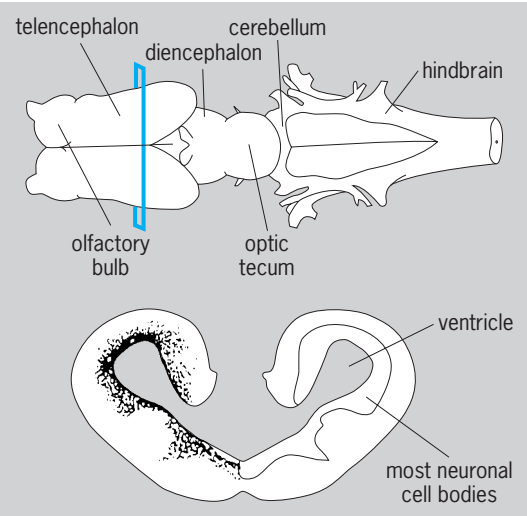


Fig. 1. Forebrain of a salamander, with a laminar brain organization. The color plane provides the cross section below. Most of the neurons remain congregated around the surface of the brain ventricle near the zone where they were "born." The cerebral hemisphere also shows invagination (rolling inward toward the midline during development). A membrane surrounding the brain (not shown) keeps the cerebrospinal fluid of the ventricles contained.

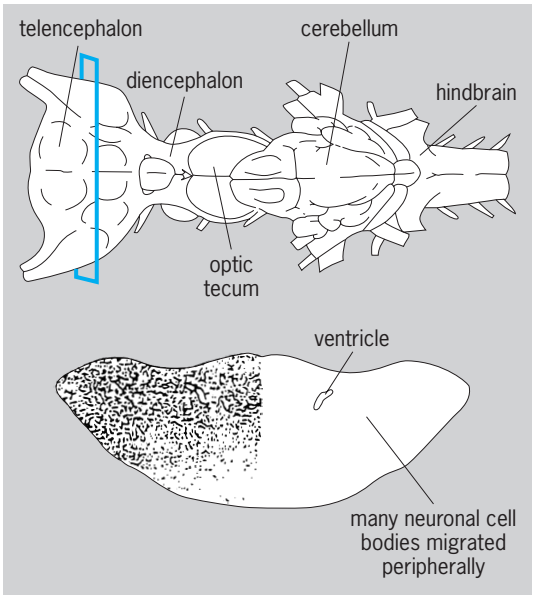


Fig. 2. Forebrain of a shark, with an elaborated brain organization. Most of the neurons have migrated away from the ventricular surface and fill virtually the entire volume of the cerebral hemisphere.

An interesting feature of the brains of bony fishes results from a developmental difference from other vertebrates. Before the embryonic nervous system rolls up to form a hollow tube, it is a relatively flat plate. In other vertebrates, the part of the plate that will form the brain rolls inward as does the rest of the nervous system to form the hollow tube, a process called invagination. In bony fishes, however, the region destined to become the brain rolls outward and downward, a process called eversion or turning outward (**Fig. 3**).

Brain size. In general, brain size or weight varies in proportion to the size of the body. In some species, however, brain size or weight is greater than would be expected for that body weight. Examples of bigger-than-expected brains are found in humans, chimpanzees, and porpoises. Although an elephant has a much larger brain than a human, it is no greater than would be expected for an animal of that body weight. Other mammals with brain weights typical for their body size are wolves, rats, and gorillas. In addition, birds have brains that are comparable to those of mammals of equivalent body weight. Some birds, such as crows, have brain weights in excess of what would be expected for their body weight. Moreover, crows have brain weights that would be expected of a small primate of equivalent body weight. In contrast to birds and mammals, amphibians, reptiles, and bony fishes have relatively small brains for their body weights, as do jawless fishes (**Fig. 4**).

Subdivisions of brain. The brain is divided into a hindbrain, a midbrain, and a forebrain.

Hindbrain. The hindbrain is a region that contains the nerve endings that receive information from the outside world and from the body interior; these are known as sensory cranial nerves. The neuron groups

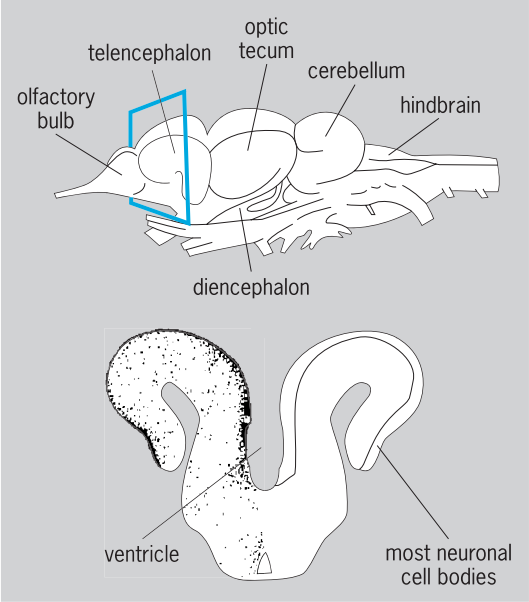


Fig. 3. Forebrain of a fish that shows eversion of the cerebral hemisphere (rolling away from the midline during development). The color plane provides the cross section below. This fish also has a laminar brain. The eversion of the cerebrum and the congregation of neurons near the ventricular surface result in the "inside-out" organization of such brains. A membrane surrounding the brain (not shown) keeps the cerebrospinal fluid of the ventricles contained.

upon which they terminate are known as sensory cranial nerves. Also found in the hindbrain are the motor nerves that control the internal and skeletal muscles and glands, which are called motor cranial nerves; the neuron groups from which they originate are known as motor cranial nerve nuclei.

Some fishes have an extraordinary sense of taste, and this is reflected in the development of their hindbrains. Such fishes are the carps and the catfishes, which are both bottom feeders that live in murky waters. In addition to having many taste buds in their mouths and throats, these fishes have thousands of taste buds studding their entire body surface. In effect, their bodies are like giant tongues that taste the water as the fish swims through it. The so-called whiskers of the catfishes, known as barbels, also are covered with large numbers of taste buds. A catfish will drag its barbels across the bottom to taste for morsels of food that might be present. Three cranial nerves are involved in the detection of taste stimuli: the facial nerve, the glossopharyngeal nerve, and the vagus nerve. The facial nerve, which innervates the skin taste buds, terminates in a large lobe on the surface of the hindbrain known as the facial lobe. The vagus and glossopharyngeal nerves, which innervate taste buds of the mouth and throat, terminate in an enormous lobe just to the side of the facial lobe, called the vagal lobe. Only carps and catfishes have these taste lobes on the hindbrain.

Many animals possess senses that humans do not possess. One such sense is the lateral line sense which derives from receptors located in the lateral line organ and can easily be seen on most bony fishes as a thin, horizontal line running the length of the body from behind the gill opening to the tail. Other lateral line organs can be found on the head and jaws. These lateral line systems contain mechanoreceptors that respond to low-frequency pressure waves that might be produced by other fishes nearby or the bow wave of a fast-swimming predator about to strike. The cranial nerves that innervate the lateral line organs terminate in a special lateral line region of the hindbrain. The lateral line sense is an effective way for animals to find out something about their world without the use of vision. Lateral line systems and a special region of the hindbrain dedicated to lateral line sense are found in fishes and sharks, jawless fishes, and bony fishes of various sorts. They are found in amphibians only in the tadpole phase of their lives and are totally lacking in birds, reptiles, and mammals. *See* LATERAL LINE SYSTEM.

Electroreception is another way of dealing with a murky environment. Scientists have described two types of electroreception: active and passive. The receptors for this sense are also located in the lateral line canals and sometimes on the skin as well. Animals with passive electroreception, such as sharks and rays, platypuses, and echidna, can detect the presence of the very weak electric fields that are generated around a living body, which they then follow to capture their prey. Animals with active

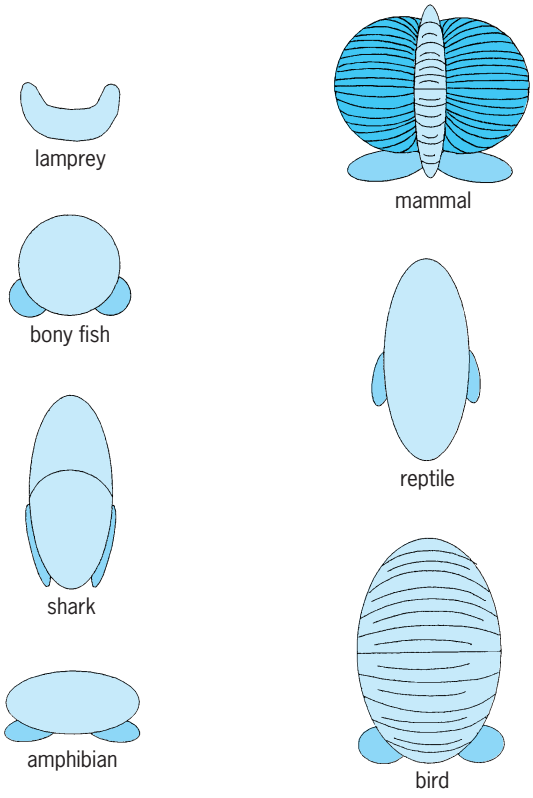


Fig. 4. Some examples of the varieties of cerebellar size and shape among vertebrates. (*Adapted from A. B. Butler and W. Hodos, Comparative Vertebrate Neuroanatomy: Adaptation and Evolution, 1966*)

electroreception generate stronger electric fields around themselves using specialized organs called electric organs. By detecting changes in the electric fields around themselves, they can derive a picture of their world including the objects and creatures in it. Electrosensory cranial nerves terminate in a region of the hindbrain known as the electrosensory area, which is close to the lateral line area. A second group of active electrosensory fishes are capable of generating electric fields that are so powerful they can stun a prey or an enemy. Among these are the electric eel, the electric catfish, and an electric shark (the torpedo). These animals also use their low-level electric fields to detect objects and creatures in the environment. *See* ELECTRIC ORGAN (BIOLOGY).

Not only did the hindbrain change in response to the great alterations in sensory evolution, but it also underwent major transformations; for example, motor-neuron groups involved in swallowing, chewing, and salivating evolved as a consequence of the transition to land and the loss of the water column to carry food from the opening of the mouth into the throat.

In addition to the sensory and motor cranial nerve nuclei, the hindbrain contains two important coordinating or integrating systems: the cerebellum and the reticular formation. The functions of the cerebellum are varied; they include the integration of a sense of balance with aspects of movement and motor learning and motor memory, as well as playing an important role in electrosensory reception. Although the internal organization of the cerebellum's neurons has remained rather conservative in evolution, the size and overall form of the cerebellum vary widely. One view is that the size of the cerebellum depends on the extent to which an animal moves through three dimensions of space rather than remaining more or less on one plane. Thus, bottom-dwelling fishes and surface-dwelling mammals would have relatively small cerebella, whereas fishes that change depth extensively and mammals that climb trees and descend to the ground, as well as flying birds and mammals, would have larger cerebella. There are, however, numerous exceptions to this rule.

The reticular formation coordinates the functions of various muscle groups that must cooperate with each other. For example, the actions of the jaws and tongue must be coordinated so that an animal does not eat its own tongue while eating its meal. The reticular formation also coordinates the motor-neuron groups that control the air column that enters and leaves the mouth and throat, which produces the various vocalizations of land animals, including speech. The reticular formation also is involved in sleep, wakefulness, and attention. In general, the reticular formation has been very conservative in its evolution, although it is best developed in reptiles, birds, and mammals.

Midbrain. The midbrain contains the motor cranial nerves that move the eyes. It also is sometimes regarded as the "map room" of the brain because it contains neuron groups that are organized to form maps of visual space, auditory space, and the body. These maps are coordinate with each other so that a sudden unexpected sound will cause the head and eyes to move to the precise region of space from which the sound originated. In those animals that make extensive use of sound localization to find their way in the world, such as owls and bats, the map areas of the midbrain are very highly developed. In addition, certain snakes, such as rattlesnakes and boa constrictors, have infrared detectors on the snout or under the eyes. These detectors can sense the minute heat from a small animal's body at a distance of 1 m (3 ft) or more. The midbrains of these animals also have infrared maps that are in register with the auditory, visual, and body maps to permit the animal to correlate all the necessary information to make a successful strike on prey in virtually total darkness.

Forebrain. The forebrain is a very complex region that consists of the thalamus, the hypothalamus, the epithalamus, and the cerebrum or telencephalon. In addition, the forebrain contains the limbic system, which has components in all regions of the forebrain as well as continuing into the midbrain. The thalamus processes and regulates a large quantity of the information that enters and emanates from the forebrain. As the cerebrum increases in size and complexity in land animals, especially in mammals and birds, the thalamus increases accordingly. The hypothalamus regulates autonomic functions, such as respiration, heart rate, blood pressure, and temperature, as well as behaviors such as feeding, drinking, courtship and reproduction, parental, territoriality, and emotional, which it controls in conjunction with the limbic system. The hypothalamus also regulates the endocrine system. The size and complexity of the hypothalamus, relative to the rest of the brain, is greatest in fishes and sharks; it declines considerably in proportion to the rest of the brain in land animals. The epithalamus contains the pineal gland, which is involved in various biological rhythms that depend on daylight, including seasonal changes such the beginning and end of the breeding season and annual migrations. In some animals, such as certain reptiles, the pineal takes on the form of an eye, located on the top of the head and known as the parietal eye. This eye has a lens and a primitive retina that capture light and transmit information, such as the amount of daylight, to the hypothalamus. The epithalamus, like the hypothalamus, is relatively smaller in the brains of land animals.

The greatest evolutionary expansion of the forebrain is seen in the cerebrum. The cerebrum consists of an outer layer, the pallium, and a series of deep structures, known as the subpallium. The subpallium is composed of the corpus striatum, the amygdala, and the hippocampus. The outer layer of the cerebrum in mammals is known as the cerebral cortex. Considerable debate surrounds the evolutionary relationship between the cerebral cortex of mammals and the pallium and subpallium of nonmammals. Most specialists, however, seem to agree that the mammalian cortex arose from the pallium and

certain regions of the subpallium. The cerebrum is relatively small in animals with laminar brains and larger in those with complex brains. Scientists have only begun to catalog the many complicated behavioral functions of the cerebrum. Among them appear to be memory, thinking and reasoning, and planning. With the advent of life on the land, the cerebrum underwent an extreme degree of elaboration in reptiles and birds and especially in mammals. *See* ENDOCRINE SYSTEM (VERTEBRATE).

William Hodos

Bibliography. M. J. Benton, *Vertebrate Paleobiology*, 1990; A. B. Butler and W. Hodos, *Comparative Vertebrate Neuroanatomy: Adaption and Evolution*, 1996; R. Nieuwenhuys, H. J. ten Donkelaar, and C. Nicholson, *The Central Nervous System of Vertebrates*, 1997.

Vertical takeoff and landing (VTOL)

A flight technique in which an aircraft rises directly into the air and settles vertically onto the ground. Such aircraft do not need runways but can operate from a small pad or, in some cases, from an unprepared site. The helicopter was the first aircraft that could hover and take off and land vertically, and is now the most widely used VTOL concept. *See* HELICOPTER.

The helicopter is ideally suited for hovering flight, but in cruise its rotor must move essentially edgewise through the air, causing vibration, high drag, and large power losses. The aerodynamic efficiency of the helicopter in cruising flight is only about one-quarter of that of a good conventional airplane. The success of the helicopter in spite of these deficiencies started a wide-ranging study of aircraft concepts that could take off like a helicopter and cruise like an airplane. The term VTOL is usually used to designate the aircraft other than the helicopter that can take off and land vertically. The term V/STOL indicates an aircraft that can take off vertically when necessary, but can also use a short running takeoff, when space is available, to lift a greater load. *See* SHORT TAKEOFF AND LANDING (STOL).

Many of the VTOL concepts of the early 1950s were carried to flight test by using small, relatively inexpensive configurations to explore flying characteristics in vertical takeoff and landing and in the transition to and from conventional flight. These programs determined the design techniques needed to cope with the unique aerodynamic, stability and control, and structural response problems of VTOL operation. By the 1960s the technology had progressed to the point that more sophisticated versions of some concepts could be considered for operational test and evaluation.

VTOL concepts may be considered under two categories, advanced rotorcraft and fixed wing. The first category includes concepts featuring rotors for vertical lift similar to the helicopter, and the second generally has a fixed wing with deflected jets.

Advanced rotorcraft. This category is subdivided into two concepts: aircraft that operate like a helicopter, and those that operate like a helicopter in hover flight but tilt the rotors to act as propellers in cruise.

Fig. 1. Bell/Boeing V-22 Osprey.

Tiltrotor. The tiltrotor concept began with the Bell XV-3 in 1958. This concept is closest to the conventional helicopter and relies heavily on helicopter technology. The rotor disks are horizontal in VTOL operation and are tilted 90° to act as propellers in cruising flight. Such aircraft can have cruise efficiencies at least twice those of the helicopter, making them especially useful for helicopter missions where greater range, speed, and time on station are desired. In 1977 the first Bell XV-15 was flown in a successful test program that further matured tiltrotor technology and led to the Bell/Boeing V-22 Osprey (**Fig. 1**), the first operational tiltrotor aircraft.

The V-22, which first flew in 1989 and is about three times the size of the XV-15, has many possible military applications, including troop carrying in Marine assaults, Army medical evacuation, Navy antisubmarine warfare, Air Force special operation forces, and Coast Guard search and rescue. A civil version of this concept could help relieve airport ground and air congestion by flying from vertiports, areas that could be made more convenient, more strategically located, and smaller than airports, so that airports could be bypassed entirely for shorter-range flights.

X-wing. Another concept explored extensively during the late 1980s was the X-wing. For vertical flight, the four rotating helicopter-type blades featured blowing along the trailing and leading edges to increase hover performance. During cruise, the rotor stops and the aircraft flies in a fixed-wing mode, with two forward and two aft swept wings. The concept went through ground-based testing, and a flight research vehicle was built. Budget constraints curtailed the program before the X-wing configuration could be flight-tested.

Other concepts. Numerous advanced rotorcraft concepts are continuously being investigated in research programs (**Fig. 2**). These include designs that take off and land as a rotor system and fix the blades to act as wings for cruise flight. Other concepts feature tiltrotor technology for vertical flight, but fold the blades and employ other propulsion modes for cruise.

Fixed wing. Development of the vectored-thrust concept began in the late 1950s. The concept features an engine with four rotating nozzles that deflect

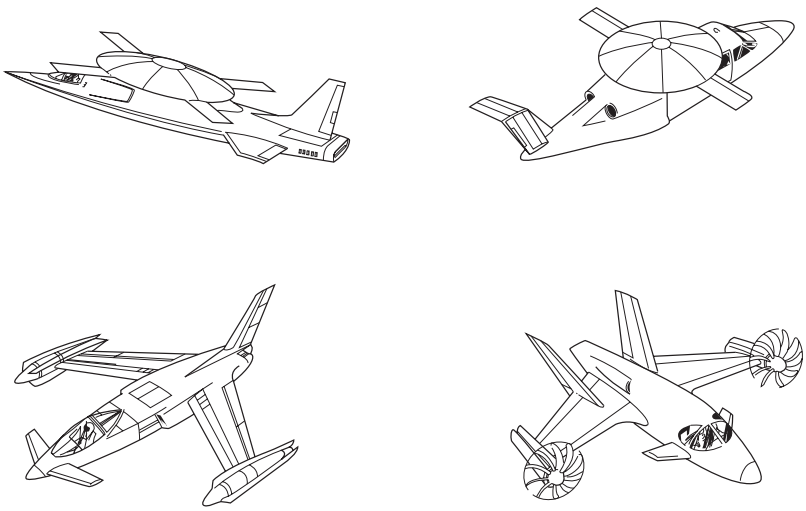


Fig. 2. Examples of advanced rotorcraft concepts.

the thrust from horizontal for conventional flight to vertical for VTOL operation. This activity led to the very successful British Harrier aircraft, which is considered to be a V/STOL aircraft because its normal mode of operation is to use a short takeoff run, when space is available, to greatly increase its payload and range capability.

The Soviet Union developed the world's second operational VTOL aircraft, the Yak 36. The aircraft features a lift-plus-lift cruise propulsion concept whereby the cruise engine has two fully vectorable nozzles for vertical and cruise flight and a pair of forward-located lift engines for lift and balance in vertical flight.

The success of the British Harrier led to the development of the Harrier II. This aircraft (**Fig. 3**) has numerous changes that provide approximately twice the range and payload of the original Harrier. The Harrier II (designated AV-8B by the U.S. Marine Corps) operates from both land and sea bases. In an amphibious operation, these aircraft fly attack and close-support missions from assault ships in support of troop-carrying helicopters. Once ashore, they are staged forward to austere sites close to the battle-

Fig. 3. McDonnell Douglas/British Aerospace Harrier II aircraft.

front, so that they can react quickly in support of the ground troops.

The initial Harrier II was powered by an engine with a thrust of 21,450 lb (95,410 newtons), but in 1989 a new engine with an additional 3000 lb (13,344 N) of thrust became an option.

With the success of the Harrier, design studies and technology development programs were directed at expanding the flight envelope of this class of aircraft to include supersonic capability. The term STOVL (short takeoff vertical landing) is used to define this supersonic fighter-attack aircraft since the large benefits in payload and range of a short takeoff will be factored into the basic design. *See* CONVERTIPLANE.

W. P. Nelms

Bibliography. J. Albers, *NASA Rotorcraft Technology for the 21st Century*, AIAA Pap. 89-2066, 1989; J. W. Fozard, *The British Aerospace Harrier: Case Study in Aircraft Design*, AIAA Professional Study Series, 1978; *Proceedings of the International Powered Lift Conference*, Santa Clara, California, December 7–10, 1987; *Proceedings of the National Technical Specialists Meeting on Tactical V-STOL Aircraft*, New Bern, North Carolina, September 19–21, 1989; *Special Course on V/STOL Aerodynamics*, AGARD Rep. 710, 1984.

Vessel traffic service

A marine traffic management system, operated by a competent authority, for improving the safety and efficiency of vessel traffic and protecting the environment within designated service areas. A vessel traffic service (VTS) has the ability and the authority to interact with marine traffic and respond to developing traffic situations. It improves order and predictability on the waterway throughout its service area by using surveillance, communications, and standard operating procedures to collect, evaluate, and share information to support professional mariners' navigation decision making. Opportunities for human error are reduced through the provision of traffic information, traffic organization, and navigation-assistance services. Captains, mates, and marine pilots use the delivered vessel traffic services to direct and control the maneuvering of ships, tugs with tows, and passenger ferries.

Systems. A modern VTS system uses a variety of sensors to collect information and assemble a "traffic image" of all vessel movements, events, and conditions in its area of responsibility. These sensors may include surface radar, closed-circuit television cameras, wind speed and visibility monitors, tide and current gauges, and simple visual overlooks. The newest tool available to vessel traffic service is the automatic identification system (AIS). This consists of a shipborne device that automatically reports each ship's position, identification, destination, and other information needed for the shore-based authority to conduct traffic management. All information is gathered from these sensors and sent to the vessel traffic center using microwave or high-capacity communications lines. There, it is presented to the vessel traffic

service operators on a sophisticated, composite display so they may monitor the complete traffic situation and relay relevant information to the participating mariners. Information is typically sent to the mariner using dedicated voice radio channels. However, more vessel traffic services are coming to rely on data transfer using digitized messages instead of ship-to-shore voice communications. Information of general interest to all vessels may be sent at regular broadcast intervals. Information or advice directed at a specific ship is sent when the ship requests information or when the vessel traffic service feels the ship needs it. *See* CLOSED-CIRCUIT TELEVISION; MICROWAVE; PILOTING; RADAR.

Operations. Functions and capabilities vary among the world's more than 300 VTS systems, but their basic operational function is to provide three services: information, navigational assistance, and traffic organization.

Information service. Information service provides ships with essential information in time to contribute to the mariner's on-board navigational decision making. This may include the position, identification, or intentions of other ships, the state of the tide or current, the availability of space in an anchorage, the presence of a concentration of fishing or recreational vessels, or the vertical clearance under a bridge. It is typically information that the vessel traffic service has gathered and that is otherwise unavailable to ships.

Navigational assistance service. Navigational assistance service is provided to an individual ship when requested, or when deemed necessary to prevent an imminent disaster. The vessel traffic service provides appropriate advice to assist with on-board navigational decision making and monitors the navigation of the vessel, but never gives specific helm or engine orders. Instead, navigation assistance provides information on a vessel's ability to adhere to its intended course or route, information on its position relative to known locations or coordinates (longitude/latitude), or the distance and direction to landmarks or other objects. Navigation assistance may include advice relating to tracks or routes to be followed, or in an extreme case directions on action to take to prevent collisions and groundings. *See* MARINE NAVIGATION; NAVIGATION; SHIP ROUTING.

Traffic organization service. Traffic organization service prevents the development of dangerous maritime situations and provides for the safe and efficient navigation of traffic within the vessel traffic service area by managing vessel movements. It concerns the forward planning of movements and water-space management, and is particularly relevant in times of congestion or when the movement of special transports may affect the flow of other traffic. Monitoring traffic, enforcing adherence to governing rules and regulations, and communicating conditions and expectations to affected parties are essential elements of traffic organization. Traffic organization may include establishing and operating a system of traffic clearances in relation to the priority of movements, allocating space, and mandatory reporting of movements, as well as establishing routes to be followed, speed limits to be observed, or other appropriate measures considered necessary by the VTS authority. It may also include managing one-way traffic routes, isolating dangerous transits, scheduling movements, or prioritizing movements. Traffic organization is used to prevent vessels from meeting or overtaking each other or to ensure that vessel interactions occur at favorable locations. Where the vessel traffic service is authorized to issue traffic organization-related instructions to vessels, the instructions given are results oriented, leaving the details of execution to the vessel. The successful operation of a traffic organization regime requires established rules, good communications, positive identification of all affected vessels, and precision positioning.

A vessel traffic service also provides support to allied services including customs and law enforcement, firefighting and emergency responders, and commercial maritime interests in the port. This support is usually in the form of information on the position, identity, or intentions of vessels of interest.

Operators. In the United States, the Coast Guard has been named in law as the competent authority to operate vessel traffic services. As such, the vessel traffic service is an extension of the Coast Guard's authority to regulate marine traffic and direct the movement of vessels. This arrangement is consistent with other VTS operations throughout the world.

Exercising this considerable authority requires a dedicated, highly trained staff of professional VTS operators. Most VTS operators come to the profession after having served at sea for many years. They must first undergo extensive training in laws and regulations, local geography and operating practices, traffic monitoring, effective communications, and equipment use. Training often includes many hours of practice using simulated traffic situations. Once certified to perform the duties of a vessel traffic service operator, they are assigned a sector and a watch shift. Most vessel traffic services operate 24 h a day throughout the year.

Users. Vessel traffic services are aimed at assisting the professional mariner. This limits participation vessels of a certain size or of a certain employment. In most areas around the world, participation in a vessel traffic service is mandatory for commercial vessels greater than 270 metric tons (300 tons) or greater than 20 m (65 ft) in length. Passenger ferries are usually required to participate. Inland or harbor craft, such as tugs, may be excluded. Small fishing vessels and recreational craft rarely take part in the vessel traffic service. In some cases, these types of craft may be prohibited from participating to reduce the communications load on the system. In other areas, these craft may be allowed to participate as needed.

Routine. The process of participating in a vessel traffic service begins when a vessel checks in to inform the vessel traffic service of its position, identity, and intentions. The medium for communicating this information is usually VHF-FM voice radio. In its initial report to the vessel traffic service, a vessel will provide the vessel traffic service with a sailing plan that will serve as the vessel traffic service's

primary source of information to monitor the participating vessel and upon which it will base its traffic organization decisions. *See* RADIO BROADCASTING.

During the course of the vessel's transit through the area, the vessel traffic service will monitor its progress and adherence to the sailing plan. The vessel traffic service may occasionally provide information to the vessel relevant to opposing traffic, channel (fairway) conditions, weather, aids to navigation, or anything else that might affect the transit or influence the sailing plan. The vessel may wish to contact the vessel traffic service during its transit for updated information or to report that it must deviate from the agreed-upon sailing plan. Upon completing its transit, a vessel might make a final call to the vessel traffic service reporting "all secure" or to inform the vessel traffic service of its plans. J. Michael Sollosi

Bibliography. International Association of Marine Aids to Navigation and Lighthouse Authorities (IALA), *Vessel Traffic Services Manual*, 1998; International Maritime Organization Assembly Resolution A.857(20), *Guidelines on Vessel Traffic Services*, 1997; National Research Council, *Vessel Navigation and Traffic for Safe and Efficient Ports and Waterways: Interim Report*, National Academy Press, Washington, DC, 1996; U.S. Coast Guard Office of Vessel Traffic Management National Research Council, *Minding the Helm: Marine Navigation and Piloting*, 1994.

Vestibular system

The vestibular system subserves the bodily functions of balance and equilibrium. It accomplishes this by assessing head and body movement and position in space, generating a neural code representing this information, and distributing this code to appropriate sites located throughout the central nervous system. The centrifugal flow of information begins at sensory hair cells located within the peripheral vestibular labyrinth. These hair cells synapse chemically with primary vestibular afferent nerve fibers, causing them to fire with a frequency code of action potentials that include the parameters of head motion and position. These vestibular afferents, in turn, enter the brain and terminate within the vestibular nuclei and cerebellum. Information carried by the firing patterns of these afferents is combined within these central structures with incoming sensory information from the visual, somatosensory, cognitive, and visceral systems to compute a central representation of head and body position in space. This representation is called the gravito-inortial vector and is an important quantity that the central nervous system employs to achieve balance and equilibrium. *See* POSTURAL EQUILIBRIUM; BRAIN; NERVOUS SYSTEM (VERTEBRATE).

Vestibular function is largely reflex and unconscious in nature. We are thus often unaware of vestibular action. For example, while we are able to remain upright in complete darkness partially because of vestibular postural reflexes, and can be aware of our posture, the mechanisms responsible for this are unconscious or unrealized. Vestibular reflexes are also analogous to spinal reflexes, like the knee jerk reflex, because, there are only a few neurons or brain cells interposed in the pathway between sensory input and motor output. These vestibular reflexes originate within the vestibular labyrinth, which is subsequently connected with the eyes (vestibulo-ocular reflexes), the neck (vestibulocolic reflexes), and the body (vestibulo-spinal reflexes). The vestibular system can influence movement, somatosensory and visual sensation, digestion, equilibrium, and state of mind. Vestibular activity modifies the firing of intrinsic cerebellar neurons, and influences spinal and brainstem motor and interneurons and superior collicular and cerebral cortical neurons, where, for example, the orientations of visual receptive fields are modified by changes in head position. *See* REFLEX.

The head can be envisioned as a sphere free to pivot upon the neck joints, which are analogous to ball and-socket joints. Thus, the head can move in three dimensions, and the appropriate labyrinthine sensors report signals that are measures of the angular velocity of the head in these three dimensions—pitch, roll, and yaw. Vertebrates also experience linear motion, for example, walking along a flat surface or climbing up and down, as well as the influence of gravity. The two components of linear acceleration—the linear acceleration vector that the brain needs to keep informed about—are the impulsive and gravitational components. Linear motion can be vertical or horizontal, and impulsive (during the initiation of linear motion) or static (gravitational). The labyrinth is thus divided into receptor areas that primarily detect angular and linear head motion.

Morphology of the vestibular labyrinth. The vestibular labyrinth is housed within the petrous portion of the temporal bone of the skull along with the cochlea, the organ of hearing (**Fig. 1**). [Anatomists named this bone the temporal bone, meaning the time bone, because this is where gray hair first appears, indicating the passage of time. Petrous refers to rock, and the petrous portion of the temporal bone is among the most dense and hard bone in the body, offering protection to the delicate cochlea and labyrinthine organs.] The receptor element or primary motion sensor within the labyrinth is the hair cell (**Fig. 2**). Hair cells respond to bending of their apical sensory hairs by changing the electrical potential across their cell membranes. These changes are called receptor potentials, and the apical surface of the hair cell thus functions as a mechanical-to-electrical transducer. Because hair cells are small with little surface area, they are isopotential throughout, and the apical receptor potentials are also sensed by the base of the cell. The cell base is a secretory organ that regulates the liberation of neurotransmitter in response to changes in the membrane potential. This neurotransmitter, probably glutamate, diffuses across the synaptic cleft to complex with chemical receptors on the peripheral or dendritic portions of fibers of the VIIIth cranial (vestibulocochlear) nerve (**Fig. 3**). The action of these chemical receptors, in turn, results in a depolarization of the membrane

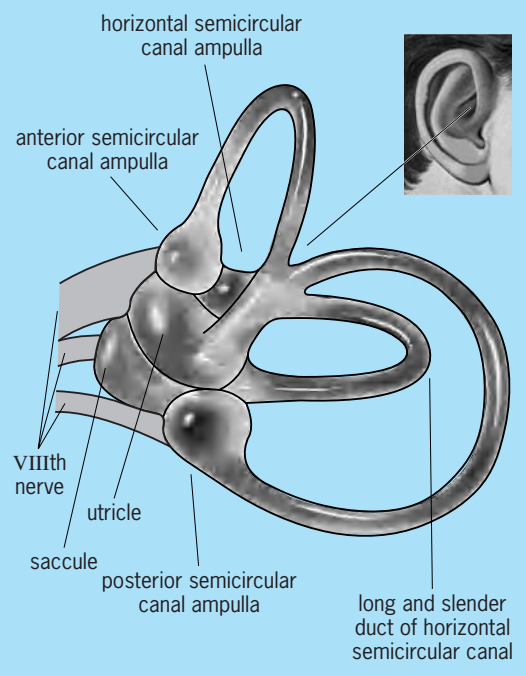


Fig. 1. The vestibular labyrinth is located within the inner ear. It communicates with the brain via the VIIIth nerve. Each of the three semicircular canals has an ampulla and a long and slender duct. The utricle primarily senses motion in an earth-parallel plane, while the saccule primarily senses motion and gravity in an earth-perpendicular plane.

of the VIIIth nerve fibers, causing the initiation of action potentials that are conducted toward the central nervous system. The frequency of this action potential train encodes the parameters of angular and linear motion. *See* BIOPOTENTIALS AND IONIC CURRENTS; EAR (VERTEBRATE); SYNAPTIC TRANSMISSION.

Hair cells are the common sensory element in both the angular and linear labyrinthine sensors as well as within the cochlea. The particular frequency of energy that hair cells sense within these diverse end organs arises because of the accessory structures surrounding the hair cells. Thus, angular motion is sensed by the semicircular canals, linear motion by the otolith organs, and sound energy by the cochlea.

Semicircular canals. The angular accelerometers, the semicircular canals, bilaterally three in number, are named for their orientation within the skull. There are two horizontal or lateral canals, two superior or anterior canals, and two inferior or posterior semicircular canals. These three canals on each side of the head are oriented in planes orthogonal to each other (Fig. 1). The three planes of space (x, y, and z or yaw, pitch, and roll) are represented by the three planes of the semicircular canals. The hair cells of each semicircular canal lie in a single sheet atop a saddle-shaped structure called the crista of the canal. The crista resides within a widened, egg-shaped portion of the canal called the ampulla (Fig. 2). Hair cells are divided into type I and II with reference to their innervation and cytoarchitectural features. The type I cell is flask-shaped with a constricted neck and a flared apical lip. The type II cell is more cylindrical. The apical surface of each cell faces a fluid with a high potassium concentration called endolymph and is covered with sensory hairs, or cilia, separated into stereocilia and a kinocilium. In the semicircular canals, a gelatinous cupula sits atop the hair cells, completely filling the lumen of the ampulla of the canal and closing it off to fluid flow. The general shape of the cupula conforms to the receptor area from which it arises. The cupula forms a diaphragm across the ampulla. It is most easily displaced at its center, where it is thinner. The hair cell kinocilium and the taller stereocilia project into the cupula and are displaced when the cupula moves. Because the cupula conforms to the shape of the crista, is attached to the ampullary wall on three sides, and is thinner in its central portions, the cilia in its central regions are subject to more displacement than those in the periphery for a given acceleration. A graded response to cupular displacement thus results. The apex of the cupula may act as a relief valve for

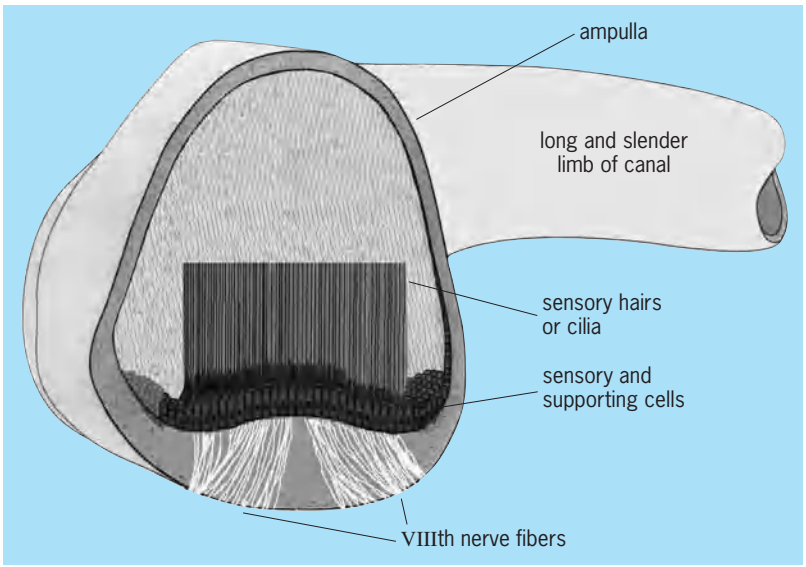


Fig. 2. Ampulla of a semicircular canal. This cross section is perpendicular to the long axis of the horizontal canal. The canal hair cell stereocilia are considerably longer than their macular counterparts, averaging about 0.1 mm in length.

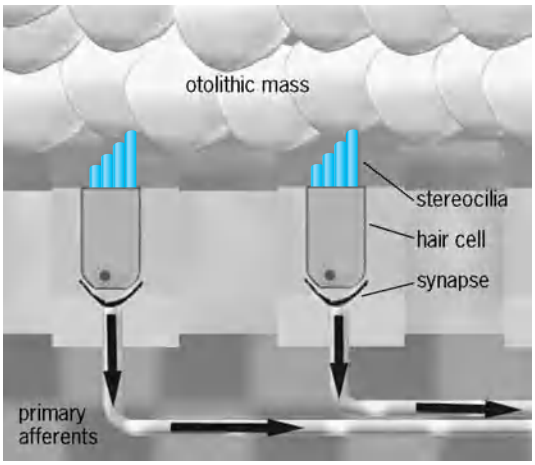


Fig. 3. Otolithic macula at rest. The arrows within the primary afferents or VIIIth nerve fibers indicate spontaneous activity in these fibers in the absence of motion of the otolithic mass relative to the hair cell stereocilia. (The otolithic membrane is not illustrated for clarity.)

extreme angular accelerational forces such as those caused by head trauma. The apical cupular surface may be displaced and ultimately detached only to reseal later. This mechanism may serve to prevent damage to receptor structures.

Otolith organs. The linear accelerometers, the otolithic organs (utricle and saccule), are bilaterally two in number. They are fluid-filled sacs containing a curved sheet of hair cells called the macula. The hair cell apical sensory hairs protrude through perforations in a membrane called the otolithic membrane to contact a mass of calcium carbonate crystals called the otolith (or stone) that is supported by the otolithic membrane (Fig. 3). The orientation of the utricular macula is roughly horizontal, and that of the saccule, roughly vertical.

Type I hair cells lie predominantly in the center of the crista and are concentrated in the striola of the macula. (The striola is a line running roughly down the center of the macula.) Type II cells are located peripherally in the crista and throughout the macula of the saccule and utricle. Type I hair cells are unique in that the afferent dendritic process of the bipolar ganglion neuron located in Scarpa's ganglion expands to form a calyx or chalice that completely surrounds the basolateral surface of the cell. Type II cells are innervated by finer fibers, and one afferent may contact up to 30–40 single hair cells. There are also the dimorphic afferents that contact both type I cells via a calyx ending and type II cells via a bouton ending. Considering the ultrastructure of the contacts between the hair cells and primary afferents, it is clear that the hair cell is the presynaptic element and the afferent the postsynaptic element within this chemical synaptic formation. Within the hair cells located at the basolateral surface opposite the afferent contact, there is almost always a typical rod or ribbon synaptic formation. In addition to the afferent synaptic processes, there are central nervous system efferent contacts arising from the efferent vestibular nuclei within the medulla and pons. These contacts terminate upon the calyx of the type I neuron and directly upon the type II hair cell and often upon its primary afferent.

The apical surfaces of vestibular hair cells are endowed with numerous cilia arranged in a characteristic fashion. The longest, the kinocilium, is eccentrically located immediately adjacent to rows of closely spaced stereocilia. The stereocilia become progressively shorter as they extend nearly to the opposite side of the cell. The hair cell is thus morphologically polarized, with the long kinocilium and tallest stereocilia at one end and the shortest stereocilia at the other end. The directional orientation of the hair cell is thus readily apparent and can be determined by noting the location of the kinocilium. The mapping of directional sensitivity has been determined for the crista and macula of the labyrinth. The least complex arrangement is found in the three semicircular canals, where the hair cells are aligned parallel to the long axes of their respective canals with the kinocilia all on the same side of the cells. In the two vertical canals, the anterior and posterior, the kinocilia are on the side opposite to the utricle, while in the horizontal canal the polarity is reversed with the kinocilia on the utricular side of the canal. In each otolithic macula the cell orientation is divided into directional hemifields separated by the striola, a line of opposed hair cells across the macular epithelium. The hair cells are oppositely oriented to each other in each hemifield.

Physiology of the vestibular labyrinth. Electrophysiological recordings have been taken from various labyrinthine elements. Elegant studies have been performed by placing microelectrodes into living hair cells, subjecting these cells to movement of their cilia, and noting changes in the membrane potential. The resting potentials of hair cells are usually between −50 and −60 mV. Hair cells respond to the movement of their cilia with transmembrane potential changes (receptor potentials) that are graded in amplitude, ranging 5–20 mV. These cells respond to static hair bundle deflections and to sinusoidal deflections with a frequency range of 0 to 150 Hz. Within this range, the output of the hair cell transduction apparatus depends on the hair bundle displacement only and not on the velocity of bundle motion (**Figs. 4** and **5**). The depolarizing phase of the receptor potential occurs upon bending of the hair bundle toward the kinocilium, and the hyperpolarizing response with bundle motion away from the kinocilium or toward the shortest stereocilium. In the semicircular canals the tallest cilia are inserted into the cupula. Estimation of the magnitude of cupular motion in the human is estimated not to exceed 3 micrometers at its extreme. Thus the sensitivity of these cells to bundle motion is extraordinary. The slope of the input/output curve can reach 20 mV per micrometer of displacement. (It has been estimated that the movement of the basilar membrane of the human cochlea is approximately equal to 1 micrometer at auditory threshold.) The depolarizing portion of the receptor potential is consistently accompanied by a decrease in input impedance of the cell.

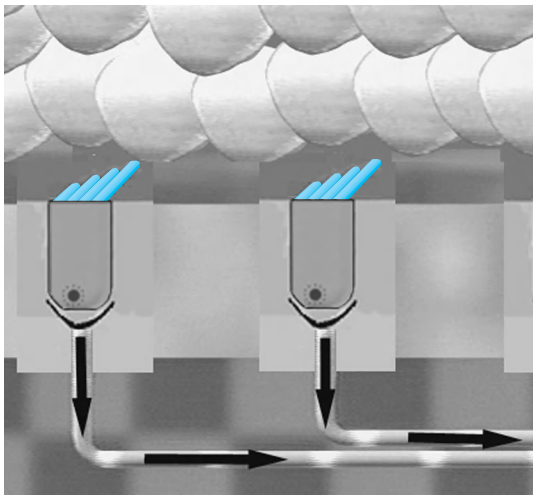


Fig. 4. Otolithic macula being stimulated by linear acceleration or gravity in its on-direction. The stereocilia are deviated toward the longest cilium (the on-direction for the hair cells). The large arrows within the primary afferents or VIIIth nerve fibers indicate increased activity in the presence of motion of the otolithic mass relative to the hair cell stereocilia.

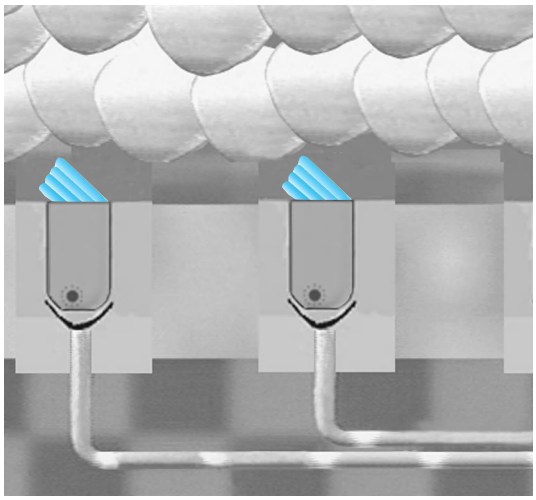


Fig. 5. Otolithic macula being stimulated by linear acceleration or gravity in its off-direction. The stereocilia are deviated away the longest cilium (the off-direction for the hair cells). The absence of arrows within the primary afferents or VIIIth nerve fibers indicates decreased or absent activity in the presence of off-direction motion of the otolithic mass relative to the hair cell stereocilia.

This suggests that the receptor potential of these vertebrate cells results from alteration of the membrane conductance to an ion whose reversal potential lies positive to the resting potential. The observation of the blocking effects of magnesium ion upon receptor potentials is consistent with a role for calcium in the transduction process. Potassium also certainly plays an important role in transduction. Thus it has been unequivocally demonstrated that hair cells can produce receptor potentials in response to the bending of their apical hairs.

The hair cell is presynaptic to the primary afferent fiber. Recording from this fiber should show modulation of any ongoing spike discharges depending upon the polarity of the receptor potential. This may be seen from single unit responses recorded within the VIIIth nerve.

Responses of the VIIIth cranial nerve. Let us consider vestibular afferents supplying the right and left horizontal semicircular canals. If an animal is placed upon a turntable and the head adjusted so that the position of the horizontal canals is parallel to the plane of rotation, rotation should modulate the receptor potentials and the horizontal canal primary afferent discharge rate. We begin rotation with a brief rapid buildup of angular acceleration, continue rotation at a constant velocity for some time, and then decelerate to a head-still position. Before stimulation begins, both horizontal canal nerves are firing with an equal resting discharge rate of around 100 impulses per second (I/S). Before considering the neural changes that might occur with stimulation, it is necessary to consider the mechanical events within the semicircular canals that result in the production of these changes. At the onset of angular acceleration the membranous labyrinth moves with the head as it is tethered to the skull. Movement of the head induces a differential pressure across the cupular membrane. This pressure causes endolymph flow relative to the canal, much as the pressure in a garden hose causes water flow out of the hose. The endolymph resists flow or movement because of its inertia, viscous drag between the fluid and canal wall, and the elastic restoring forces of the cupula and hair cell cilia. Then, during the initial angular acceleration there is a brief period of endolymph flow due to the motion of the canal and the relative stationariness of the fluid within the canal. This results in cupular deflection, say, in the utriculopedal direction in the right horizontal canal, leading to an increase in the firing rate of nerve fibers supplying the right side. Conversely, directly opposite actions take place on the left canal, effectively decreasing or even completely shutting down the firing rate of fibers supplying the left side. Moments later during the constant-velocity phase of rotation, as inertia is overcome the endolymph begins to catch up with surrounding structures, and there is no longer a relative endolymph flow acting against the cupula. Nevertheless, the displacement of the cupula continues for a brief period due to its inherent sluggishness. After a time, when all transient disturbances subside and angular velocity remains constant, the rate of firing of the nerve fibers returns to its prestimulus value. When the head is abruptly decelerated, this force acts upon the bilateral horizontal canals, causing reactions that are precisely opposite to those which occurred upon acceleration. Note that during deceleration inertia works to sustain motion in the endolymph while movement in other structures abruptly stops.

In summary, endolymph flow in a utriculopedal direction deforms the cupula, which then carries with it the cilia, causing a depolarizing receptor potential to be generated at the apex of the hair cell. This depolarization causes the basal or secretory portion of the hair cell to secrete more neurotransmitter, which increases the firing rate of the vestibular primary afferent. Reversing the direction of the endolymph flow hyperpolarizes hair cells, decreases ongoing transmitter release, and decreases the afferent firing rate to a value below the resting discharge rate. These responses are mirrored in the opposite semicircular canal. Thus acceleration and deceleration are signaled bidirectionally via a modulation of ongoing spontaneous neural activity. These actions are well correlated with the morphological polarity of hair cells in the lateral canals, where it is now known that bending of the cilia toward the kinocilium results in hair cell depolarization, whereas hyperpolarization results from bending the cilia away from the kinocilium or toward the shortest stereocilium. Just as the bilateral horizontal canals work in tandem, the right anterior and left posterior semicircular canals are roughly coplanar and work in tandem. The opposite is true for the right posterior and left anterior canals.

In the macula of the utricle or saccule, the adequate stimulus is impulsive linear acceleration or a change in the effective gravitational force caused by tilting of the head. As far as the hair cells are concerned, this is directly analogous to what happens within the crista of the semicircular canals except that the driving force, bending of the sensory hairs, is now caused by relative motion of the otolithic stone.

If the head is tilted in either pitch or roll, the otolith will tend to move relative to the head because of the effect of gravity upon its weight. This motion will impart a force on the hair bundles, causing them to bend and initiating the transduction cascade that reports this head tilt to the brain. If one experiences an impulse of linear acceleration such as beginning to walk, the inertia of the otolith will cause it to lag behind the head, thus bending the sensory hairs. When linear acceleration is complete, as in the constant-velocity phase of walking, the stone will return to its original position because of the elastic restoring forces of the otolithic membrane. Just as in the semicircular canals, these responses are bilaterally paired and signal appropriate bidirectional increases and decreases in impulsive and gravitational components of linear acceleration.

Unusual combinations of angular and linear motion can cause motion sickness or sea sickness. Terrestrial motion sickness is analogous to the space adaptation syndrome that afflicts astronauts during their sojourn into microgravity. In the case of space travel into microgravity, the weight of the otolith becomes small, nearly zero, while the mass of the otolith remains unchanged. Thus there are no responses to tilt while responses to impulsive linear accelerations persist. The semicircular canals function normally, as they have no gravitational responses. This unusual combination of impulses, that is, normal canal and impulsive linear inputs with no gravitational inputs traveling to the brain, might provoke a feeling of nausea and imbalance and confound the orientation and magnitude of the gravitoinertial vector, resulting in space adaptation syndrome or motion sickness. However, recall that there are additional sensory inputs that contribute to this vector. Astronauts report that being strapped into their couches may ameliorate their symptoms, just as sailors report that staring out at the horizon may ameliorate sea sickness.

Overview. The crista ampularis of the vestibular semicircular canals contains two types of receptor neurons or hair cells: Type I cells are flask-shaped and centrally located, while type II cells are more cylindrical and predominate in the periphery of the crista. Within the macula of the saccule and utricle, type I cells predominate in the striola where they are morphologically oppositely polarized, while type II cells are found throughout the macula. Both type I and II cells are presynaptic to axons of the vestibular portion of the VIIIth cranial nerve. Mammalian vestibular nerve fibers range in diameter from 1 to 10 micrometers and are usually grouped into two classes, thick and thin fibers. Thick fibers form a nerve calyx terminal expansion that surrounds the cell body of type I hair cells. From one to five hair cells contact these thick fibers. By contrast, the finer afferent fibers are innervated by many type II hair cells. The thick vestibular afferents discharge phasically, while the thin fibers discharge more tonically and generally have a lower sensitivity to natural stimulation. The phasic nature of the type I system is presumably due to afferent innervation by only a few hair cells. Their higher sensitivity may be ascribed to the character of the diaphragmatic displacement of the central portions of the cupula containing the cilia of the type I cells. The central portions of the cupula are thinner than the peripheral portions and undergo a greater displacement for a given head movement than the peripheral portions. The kinocilia of the type II cells located in the thicker, less mobile peripheral portions of the cupula probably endow the type II fine-fiber system with its low sensitivity to adequate stimulation.

Central projections of vestibular primary afferents. The primary afferents innervated by hair cells are the peripheral processes of bipolar neurons having cell bodies located in Scarpa's ganglion within the internal auditory meatus. The central processes of these cells contact neurons in the brainstem of the central nervous system. These centrally directed processes pass between the inferior cerebellar peduncle and the spinal trigeminal tract and bifurcate into rostrally and caudally directed fascicles which are then distributed within the vestibular nuclei and cerebellum. Thick and thin vestibular fibers terminate upon specific portions of the vestibular nuclei. A portion of the primary vestibular fibers project directly to the cerebellum via the juxtarestiform body and terminate ipsilaterally in the lingula, uvula, and nodulus. There are no direct projections to the cerebellar flocculus or paraflocculus. Second-order vestibular fibers arising from cell bodies within the vestibular nuclei also project to the cerebellum, including the flocculus and ventral paraflocculus, but do so bilaterally.

Vestibular nuclei. This complex is defined as the brainstem region where primary afferents from the labyrinth terminate. The nuclear complex is composed of four main nuclei: the superior, medial, lateral, and descending nuclei. The axonal projections of vestibular nuclear neurons travel to all parts of the neuraxis, including the brainstem, cerebellum, spinal cord, and cerebrum. Ascending fibers originating within the superior, medial, and ventrolateral nuclei travel within a fiber tract called the medial longitudinal fasciculus to terminate within the extraocular motor nuclei to subserve the vestibulo-ocular reflex. These fibers also have axon collaterals in other midbrain, pontine, and medullary nuclei that provide efference copy of the intended eye movement. The medial and ventrolateral nuclei also send fibers to the rostral extraocular motor nuclei that bifurcate, sending a branch via the descending medial longitudinal fasiculus to terminate upon cervical motoneurons and interneurons. These bifurcating fibers help to coordinate the vestibulo-ocular and vestibulocolic reflexes to stabilize eye and head in space. Many of these vestibulo-ocular, vestibulo-spinal, and vestibulo-ocular-spinal neurons receive terminals of floccular and ventral paraflocculus Purkinje neurons. In general, vestibular nucleus neurons maintain a high degree of spontaneous discharge. The motor outputs caused by the axons of vestibular neurons are thus under the modulatory control of the cerebellum. *See* MOTOR SYSTEMS.

The dorsal portion of the lateral vestibular nucleus is called Deiters' nucleus and is the origin of the ipsilateral descending vestibulo-spinal tract that terminates upon ipsilateral extensor motoneurons at several levels of the spinal cord. Dorsal Deiters' neurons are the recipient of heavy input from Purkinje cells residing within the cerebellar vermal anterior lobe. This lateral vestibulo-spinal tract and nucleus subserves the vestibulo-spinal reflexes that help maintain postural stability. The excitatory drive upon extensor motoneurons provided by the lateral vestibulo-spinal tract is counterbalanced by a similar input to flexor motoneurons that originates from the red nucleus. The descending nucleus and the caudal portions of the medial vestibular nucleus have a network of back-and-forth connections with the cerebellum. These pathways undoubtedly play a role in maintaining posture by modulating the impulse traffic that enters into motor pathways.

Efferent vestibular system. In all vertebrates, including the earliest evolved such as lamprey or myxine (hagfish), there is an efferent system that originates from cell bodies within the central nervous system and terminates upon labyrinthine hair cells and primary afferents. In primates these efferent vestibular cell bodies reside within the efferent vestibular nuclei adjacent to the abducens nuclei at the pontomedullary junction. Electrical stimulation of this efferent system causes excitation in vestibular primary afferents but inhibits hair cells. The efferent vestibular system is presently a subject of intense study but undoubtedly is in place to enhance vestibular function. It is interesting that evolution felt it necessary to modify incoming vestibular information before it could enter the central nervous system.

Stephen M. Highstein

Bibliography. P. Blazquez et al., Input of the anterior and posterior semicircular canals via interneurons carrying head velocity information to the dorsal Y group of the vestibular nuclei, *J. Neurophysiol.*, 83:2891–2904, 2000; R. Boyle et al., Neural readaptation to Earth's gravity following return from space, *J. Neurophysiol.*, 86:2118–2122, 2001; S. M. Highstein, R. D. Rabbitt, and R. Boyle, Determinants of semicircular canal response dynamics in the toadfish, *Opsanus tau*, *J. Neurophysiol.*, 75:575–596, 1996; R. A. McCrea, A. Strassman, and S. M. Highstein, Anatomical and physiological characteristics of vestibular neurons mediating the vertical vestibulo-ocular reflex of the squirrel monkey, *J. Comp. Neurol.*, 264:571–594, 1987; R. B. Silver et al., Examination of the cupula and stereocilia of the horizontal semicircular canal in the toadfish, *Opsanus tau*, *J. Comp. Neurol.*, 402:48–61, 1998.

Vestimentifera

A group of benthic marine annelid worms, discovered in 1969, that is restricted to habitats rich in sulfide (for example, hydrothermal vents and sulfide seeps). As adults they lack a mouth, gut, and anus, and are nourished by internal symbionts. All vestimentiferans live in tubes of varying hardness and rigidity (see **illustration**). Tube material is secreted by internal glands and is a mixture of chitin and protein. *See* HYDROTHERMAL VENT.

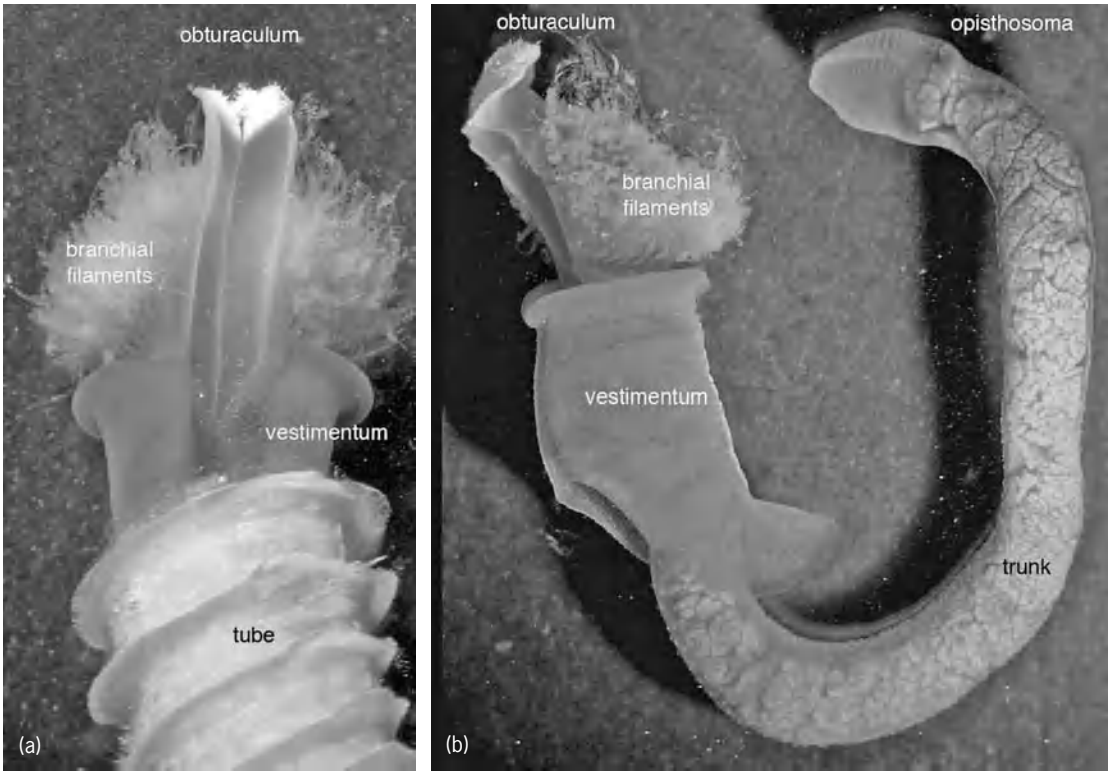
Classification. In the past, vestimentiferans had been considered to be members of the phylum Pogonophora as well as of the phylum Annelida, and it even had been suggested that they are a phylum in their own right. This was largely resulted from emphasis on their extraordinary anatomy rather than focusing on their marked similarities with tube-dwelling annelids, such as sabellid and serpulid tubeworms. They are now considered to be a part of the annelid family Siboglinidae along with other members of the former phylum Pogonophora and the bone-eating worms *Osedax*. *See* ANNELIDA.

Taxonomy. The previously erected taxonomy within Vestimentifera has been superseded following the placement of the group inside Siboglinidae. Currently there are 10 genera containing a total of 15 species: *Alaysia*, *Arcovestia*, *Escarpia*, *Ridgeia*, *Riftia*, *Lamellibrachia*, *Oasisia*, *Paraescarpia*, *Seepiophila*, and *Tevnia*. Further study is needed to establish whether other groupings apart from genera are needed.

Morphology. Vestimentiferans have four regions along their length (see illustration). The anteriormost, the obturacular region, has a mass of vascularized filaments supported by a paired structure, the obturaculum. Normally these branchial filaments protrude from the opening of the tube and act as a gill, allowing for exchange of dissolved substances between the worm's body and seawater. There may be as many as 230,000 filaments in this branchial plume.

The second region, the vestimentum, is muscular and serves to maintain the plume of branchial filaments in the open seawater by pressure on the inner tube surface at its opening. The vestimentum is the site of many glands that contribute material for lengthening the tube and thickening it near the opening. Laterally and dorsally, a pair of elongated flaps overlap one another and form a chamber that communicates with the seawater. The genital pores, both female and male, open into this chamber, and eggs and sperm pass through it.

The third region, the trunk, is a single segment and constitutes about 75% of the total length of the worm. It has a pair of large longitudinal blood vessels; the dorsal of these returns deoxygenated blood and carbon dioxide to the anterior branchial plume and, with its muscular wall, functions as a heart. The ventral vessel brings oxygenated blood from the plume to the body of the worm. The bulk of the trunk is made up of the trophosome, an organ containing masses of sulfide-oxidizing bacteria. Gonads are also present in the trunk. Much of the body wall of the trunk is lined internally by longitudinal muscles that allow the worm to withdraw into its tube for protection from predators. Glands secreting tube material add to the thickness of much of the tube posterior to the opening.



Tevnia jerichoana. **(*a*) In tube. (*b*) Removed from tube, showing four body regions.**

The fourth and most posterior region, the opisthosoma, is made up of many segments, each segment in the anterior portion bearing a row of small hooks that can be set into the inner surface of the tube. This provides an anchor against which the body of the worm retracts when the longitudinal muscles of the trunk contract during withdrawal into the tube.

Nutrition. The blood of vestimentiferans is red owing to hemoglobin, similar to that of other annelids. Oxygen and sulfide are transported by the hemoglobin, and are bound so tightly that they cannot interact as they would in normal circumstances. Thus, the symbiotic sulfide-oxidizing bacteria are able to carry out their chemoautotrophic mode of nutrition utilizing the sulfide and oxygen. The oxygen also supports the metabolism of the worm. There is no mouth, gut, or anus in postlarval worms, and their nutrition relies solely on the symbiotic bacteria contained in the bacteriocytes (specialized cells) of the trophosome. Certain cells of the worm harvest bacteria by breaking down bacteriocytes, and the food is passed on for the growth and metabolism of the worm. *See* BACTERIAL PHYSIOLOGY AND METABOLISM; HEMOGLOBIN.

Nervous system. The brain is present in the anteroventral portion of the vestimentum. The ventral nerve cord extends from the brain, separates around the ventral ciliated field, reunites, and continues posteriorly to the posterior tip of the opisthosome. Giant axons are associated with the ventral nerve cords, allowing the organism to enhance its withdrawal response into the tube. Nerve tracts pass anteriorly from the brain to each of the branchial filaments; they appear to coordinate the contractions of a series of ring muscles around the blood vessels of the branchial filaments and help move blood through the filaments back to the central region of the blood system. Other nerve tracts pass to the brain from certain sensory filaments associated with the branchial filaments.

Digestive, excretory, and reproductive systems. An excretory organ comprising a pair of nephridia, as in many other tubicolous annelids, is situated ventral and anterior to the brain. It consists of a mass of ciliated tubules that appear to unite and ultimately open into the paired excretory ducts. The ducts may further fuse to a single opening to the exterior.

A transitory digestive system has been demonstrated in one vestimentiferan species and is likely to be found across the group. Late larval stages exhibit an anterior, ventral–medial process. This snout-like structure has an opening (mouth) surrounded by cilia and leads to a continuous (intestinal) cavity proceeding to a terminal anus. When the larvae reach 400 μm long, the snout and the anus are no longer present. Only the former gut remains; however, it is isolated and contains bacteria. It is believed that when the formerly planktonic larvae settle in an area that contains high levels of sulfide, they feed on benthic bacteria, among which are sulfide-oxidizing bacteria. The sulfide-oxidizing bacteria appear to be phagocytized by cells of the midgut epithelium, where they grow and multiply. When the mouth and anus close, the midgut, with its associated bacteria, increases in size and complexity and becomes the trophosome.

Reproductive system. There are separate sexes in Vestimentifera. Mature eggs range 75–110 μm in

diameter and are the result of oogenesis in paired anterior ovaries. They pass posteriorly, then anteriorly, through paired oviducts to emerge through paired gonopores on the dorsal surface of the vestimentum.

Spermatogenesis is initiated in the testicular epithelium of tubules that open into paired sperm ducts. As many as 1500 spermatozoa may be associated with a single cytophore in the anterior portion of the sperm ducts. Near the gonopore, spermatozoa are free of the cytophore and are arranged in bundles. Spermatozoa are shed in bundles (spermatozeugmata) through paired gonopores on the dorsal surface of the vestimentum.

Larval development. Fertilization of eggs is internal in the oviducts. Sperm either transferred directly from a male or gathered from the seawater are stored in a special region of each oviduct known as a spermatheca. The eggs are fertilized as they pass through the spermathecae and are then spawned into the water and develop into trochophore larvae. Branchial filaments appear shortly thereafter. Opithosomal chaetae develop in the posterior segments. The vestimentum is formed when the larval length is about 400 μm. The obturaculum is 1 mm long. Internal bacteria are consistently present at larval lengths greater than 285 μm. *See* HYDROTHERMAL VENT.

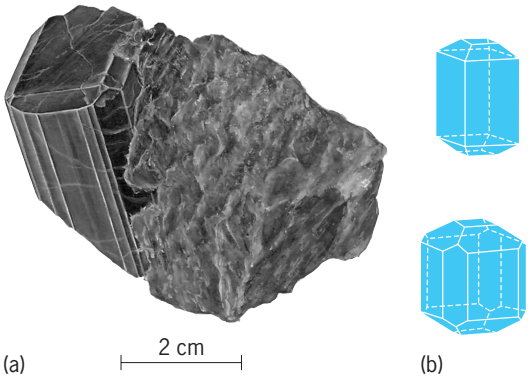
Meredith L. Jones; Greg Rouse

Bibliography. C. R. Fisher, Chemoautotrophic and methanotrophic symbioses in marine invertebrates, *Rev. Aquatic Sci.*, 2:399–436, 1990; A. Hilario, C. M. Young, and P. A. Tyler, Sperm storage, internal fertilization, and embryonic dispersal in vent and seep tubeworms (Polychaeta: Siboglinidae: Vestimentifera), *Biol. Bull.*, 208:20–28, 2005; M. L. Jones, The giant tube worms, *Oceanus*, 27:47–54, 1984; G. W. Rouse, A cladistic analysis of Siboglinidae Caullery, 1914 (Polychaeta, Annelida): Formerly the phyla Pogonophora and Vestimentifera, *Zool. J. Linn. Soc.*, 132:55–80, 2001; A. Schulze and K. M. Halanych, Siboglinid evolution shaped by habitat preference and sulfide tolerance, *Hydrobiologia*, 496:199–205, 2003; M. Webb, *Lamellibrachia barbami*, gen. nov. sp. nov. (Pogonophora), from the northeast Pacific, *Bull. Mar. Sci.*, 19:18–47, 1969.

Vesuvianite

A sorosilicate mineral of complex composition crystallizing in the tetragonal system; also known by the name idocrase. Crystals, frequently well formed, are usually prismatic with pyramidal terminations (see **illus.**). It commonly occurs in columnar aggregates but may be granular or massive. The luster is vitreous to resinous; the color is usually green or brown but may be yellow, blue, or red. Hardness is $6^1/_2$ on Mohs scale; specific gravity is 3.35–3.45.

The composition of vesuvianite is expressed by the formula $Ca_{10}Al_4(Mg,Fe)_2Si_9O_{34}(OH)_4$. Magnesium and ferrous iron are present in varying amounts, and boron or fluorine is found in some varieties. Beryllium has been reported in small amounts.



Vesuvianite. (*a*) Crystal, Christiansand, Norway (*specimen from Department of Geology, Bryn Mawr College*). (*b*) Crystal habits (*after C. Klein and C.S. Hurlbut, Jr., Manual of Mineralogy, 21st ed., 1993*)

Vesuvianite is found characteristically in crystalline limestones resulting from contact metamorphism. It is there associated with other contact minerals such as garnet, diopside, wollastonite, and tourmaline. Noted localities are Zermatt, Switzerland; Christiansand, Norway; River Vilui, Siberia; and Chiapas, Mexico. In the United States it is found in Sanford, Maine; Franklin, New Jersey; Amity, New York; and at many contact metamorphic deposits in western states. A compact green variety resembling jade is found in California and is called californite. *See* SILICATE MINERALS.

Cornelius S. Hurlbut, Jr.

Vetch

Any of a group of plants which are mostly annual and perennial legumes with weak viny stems often terminating in tendrils. There are about 150 species in the temperate zones of four continents. The leaves are compound with many leaflets. Vetches are used mainly for green manure, cover crops, hay, and pasture and silage. The seeds are used as concentrate in animal feeds, and some vetches are used as a vegetable for human consumption. Cool temperatures promote best development. Seed is sown in the fall in the Southern United States and in early spring in the Northern states. In general, are vetches more tolerant of acidic soil conditions than are most legume crops, but they have a relatively high requirement for phosphorus. Inoculation of seed with nodule-forming bacteria for symbiotic nitrogen fixation is necessary if vetch has not been previously grown in the field. *See* LEAF; LEGUME.

Traditionally vetch species have been grown in mixes with cereals for use as fodder. Although it was formerly thought that vetch species were self-fertile, recently it has been determined that some of the species are not self-fertile and actually benefit from bees or other pollinating insects for seed production. Several of the species are considered good plants for bee honey production. Identification of vetches is difficult until pods and seeds develop. *See* POLLINATION.

United States species. By the 1960s, some 35 vetch species and subspecies had been found in the United States. Many of these species have been introduced from Europe, the Far East, and Asia, where they had been used for several centuries for forage and as vegetables. Of these 35 species, 16 are native to the United States.

Vicia villosa. The vetch complex *Vicia villosa* is the most widely grown. The subspecies of this complex (*V. villosa* ssp. *villosa*), commonly known as hairy vetch, is the most winter-hardy and is mostly adapted to the eastern and southern United States, where it is grown as a winter annual or biannual. A second subspecies of this group is woollypod or smooth vetch (*V. villosa* ssp. *varia*). It is less winter-hardy than the hairy vetch type and is adapted to the Pacific Coast states, where it grows as a winter annual, and the southern United States. The first variety of woollypod vetch, Lana, was developed in 1961 in California. Recently new varieties, such as Cepello and Haymaker Plus, have been developed in Australia and are imported for use in California. Seed lots of hairy vetch that are sold often have mixtures of hairy and woollypod or smooth vetch. *Vicia villosa* ssp. *pseudocracca* occurs as a naturalized type in the northeastern United States.

Vicia sativa. The vetch complex *V. sativa* is the second most important species in use in the United States. In its native habitat of Europe and the Mediterranean regions, this vetch is known to occur as five subspecies. Common vetch (*V. sativa* ssp. *sativa*) is used throughout the United States. However, due to its it low winter hardiness, it is grown as a winter annual in the southern and western states and as a spring-summer annual in the northern states. Narrow-leaf or blackpod vetch (*V. sativa* ssp. *nigra*) occurs mostly as a weed in waste places in the United States. It spreads by volunteering and makes an excellent pasture and covercop. Underground vetch (*V. sativa* ssp. *amphicarpa*) produces pods above and below the ground and is being studied for revegetation on marginal lands of the Far East.

Vicia faba. Commonly known as the faba bean, horsebean, or broadbean, *V. faba* produces coarse, upright plants with large leaves and pods. It is one of only several vetches that are important sources of food for human beings. Some varieties are grown as vegetables. Frequently dried, faba beans have been a part of cuisines around the Mediterranean basin, and the Chinese have eaten them for more than 5000 years. It is also used in concentrate feeds for livestock in the Mediterranean basin.

This species can be divided into two subforms based on seed size. The horsebean is the large-seeded form (*V. faba* var. *faba*), and the tick or bell bean (*V. faba* var. *minuta*) is the small-seeded form, which is often used as a green manure-covercrop in the warmer regions of the United States during the winter and spring. In the colder regions, it is used during the spring and summer. In the Mediterranean basin the small-seeded form is used as a stock feed as well as a vegetable.

Fig. 1. Purple vetch in flower. (*USDA*)

Other species. Purple vetch (*V. bengalensis*) possesses poor winter hardiness; however, it is useful in the California rice-growing area, where it is used as a soil-improving crop in rotation with the rice (**Fig. 1**). Hungarian vetch (*V. pannonica*), a more winter-hardy type from Central Europe, is grown in the Pacific Northwest for forage, green manure, and seed.

One-flowered vetch (*V. articulata*) has fine leaves and stems but, lacking winter hardiness, is confined to warm regions. It can be grown under the same conditions as common vetch. Plants are smooth or nearly so, and the light-lavender-colored flowers are borne singly. Its flat seeds distinguish it from its nearest look-alike, bard or barn vetch (*V. monantha*), which is very restricted in adaptation. Bitter vetch (*V. ervilia*) is an erect vetch without tendrils. Although it is mainly found in southern Europe, it has been grown in the Pacific Northwest for use as stock feed in small quantities, for it can be toxic to livestock in larger quantities.

Bigflower or large yellow vetch (*V. grandiflora*), an introduced annual that has naturalized in the southern states, the western seaboard states, and the Great Lakes states, has promise as winter cover and reseeds freely. Woodford, a cultivar, was released in Kentucky for use on less-developed (rough) pastures.

Tiny vetch (*V. hirsuta*) is an introduced annual that has naturalized along the southern states bordering the Mississippi River, Texas, the eastern seaboard states, the Great Lakes states, and the Pacific Coast states.

Narbonne or French vetch (*V. narbonensis*), a possible wild relative of *V. fava*, is being studied in Australia, Europe, and the Middle East for use as a food, feed, and fodder crop. It has been cultivated as a forage crop in the eastern United States and has locally established as a weed in Maryland and the District of Columbia.

Weeds. Bird or purple-white tufted vetch (*V. cracca*) is a winter-hardy perennial, and its subspecies, cow or bramble vetch (*V. cracca* ssp. *tenuifolia*), is considered to be an invasive weed in the eastern United States and parts of Canada. This species complex was introduced for erosion control on cutover and burnt-over areas. Four-seeded, sparrow, or lentil vetch (*V. tetrasperma*), also considered an invasive weed, is an introduced annual that has naturalized along and east of the Mississippi River and in the Pacific states.

Crown vetch. This plant, *Coronilla varia*, is a long-lived, winter-hardy perennial legume, but it is not a true vetch. It spreads by seeds and rhizomes to form a dense, weed-free, erosion-resisting ground cover. Its greatest use is for erosion and weed control on unmowed slopes of highways, industrial developments, and military installations. Its forage value is being studied, Crown vetch has wide adaptation to variations in climate and soils. It is highly regarded for its resistance to fire. Rose, pink, and white blossoms enhance its attractiveness throughout most of the growing season (**Fig. 2**). It thrives best on well-limed and well-drained soils, requires no mowing, and rarely needs fertilization. Plantings are made with crowns (roots) or with seed that require special bacterial inoculation.

Crown vetch is slow to germinate and become established. Diseases and insects are of minor importance. Neither spittlebug nor weevil attacks crown vetch. Grazing animals do not bloat on it. Seed originally was produced only in Pennsylvania, starting in 1946; now it is grown in several states, including Ohio, New York, Nebraska, Iowa, and West Virginia. The first variety named and released was Penngift; later the Emerald (Iowa) and the Chemung (New York) varieties were named and released.

Some research has shown the palatability of crown vetch to be low except while growth is young and tender. In other grazing trials, animals were slow to accept it, but after a few days their performance on crown vetch was comparable to that on other common grass-legume pastures. Chemical analysis of crown vetch hay has shown that its crude protein and crude fiber content is similar to that of other legume hays. Digestible dry matter of crown vetch hay is below that of other grass-legume hays harvested at the same stage of maturity. Crown vetch hay is often difficult to wilt and cure. Its forage value is being studied. *See* COVER CROPS; SOIL CONSERVATION.

Fig. 2. Crown vetch. (*a*) In flower. (*b*) In seed. Each segment of a pod contains a seed.

Seed. Vetch seed is produced commercially along the Pacific Coast and in the southern Great Plains. and the South. Although most of the demand for vetch seed is supplied by production in the United States, some seed (of woollypod vetch) is imported from Australia, and seed of broad and tick beans is imported from Canada. At one time, much hairy vetch seed was imported from Europe. Vetch is grown alone and with small grains which give mechanical support to the weak stems of the vetch. Seed retains vitality for 5 years or longer. *See* SEED.

Diseases and insect pests. Diseases of vetch include anthracnose, false anthracnose gray mold, downy mildew, black stem, stem rot, root rot, rust, ovularia leaf and stem spot, several mosaic viruses, and septoria scald. Root-knot nematodes attack all varieties. Control is aided by using uninfected seed of resistant varieties on clean land. *See* NEMATA; PLANT PATHOLOGY.

Vetch is attacked by many of the insect pests of alfalfa, clover, and other forage legumes, including the pea aphid, cutworm, corn earworm, fall armyworm, vetch bruchid, grasshopper, lygus bug, and leafhopper.

Various hybrids of common vetch have been developed to be resistant to the bruchid insect and certain nematodes. *See* APHID; ENTOMOLOGY, ECONOMIC; HEMIPTERA; INSECTA.

Estimating nitrogen contribution. With the current interest in using sustainable agriculture methods to produce food in the United States, the vetches have taken on a renewed importance for those farmers who want to use legumes for crop rotation and green manure. Some of the vetch types like woollypod and hairy vetch show the ability to fix high amounts of nitrogen through their root symbiosis with nitrogen-fixing bacteria. Since these vetches and other types are being used to enrich the soil to help other crops, it is useful to be able to estimate the above nitrogen production of these vetches. In the 1950s, a method was developed by researchers at the University of Davis that is still in use today. The method involves sampling the aboveground fresh weight of the vetch crop, from which it is possible to estimate the amount of nitrogen in the sample. *See* NITROGEN FIXATION; AGRICULTURAL SOIL AND CROP PRACTICES.

Fred V. Grau; Walter Graves

Bibliography. J. A. Duke, *Handbook of Legumes of World Economic Importance*, Plenum, New York, 1981; R. R. Henson and H. A. Schoth, Vetch Culture and Uses, *USDA Farmers' Bull.*, no. 1740, 1955; F. J. Herman, Vetches in the United States, Native Naturalized and Cultivated. *USDA Agr. Handb.*, no. 168, 1960; J. H. Wiersema, J. H. Kirkbride, Jr., and C. R.

Gunn, Legume (Fabaceae) Nomenclature in the USDA Germplasm System, *USDA ARS Tech. Bull.*, no. 1757, 1990; W. A. Williams and W. L. Graves, Nitrogen contribution of annual legumes, in *Agron. Prog. Rep.*, no. 266, University of California Agricultural Experiment Station and Cooperative Extension, 1999.

Vibration

The term used to describe a continuing periodic change in the magnitude of a displacement with respect to a specified central reference. The periodic motion may range from the simple to-and-fro oscillations of a pendulum, through the more complicated vibrations of a steel plate when struck with a hammer, to the extremely complicated vibrations of large structures such as an automobile on a rough road. Vibrations are also experienced by atoms, molecules, and nuclei.

A mechanical system must possess the properties of mass and stiffness or their equivalents in order to be capable of self-supported free vibration. Stiffness implies that an alteration in the normal configuration of the system will result in a restoring force tending to return it to this configuration. Mass or inertia implies that the velocity imparted to the system in being restored to its normal configuration will cause it to overshoot this configuration. It is in consequence of the interplay of mass and stiffness that periodic vibrations in mechanical systems are possible.

Mechanical vibration. This is the term used to describe the continuing periodic motion of a solid body at any frequency. When the rate of vibration of the solid body ranges between 20 and 20,000 Hz, it may also be referred to as an acoustic vibration, for if these vibrations are transmitted to a human ear they will produce the sensation of sound. The vibration of such a solid body in contact with a fluid medium such as air or water induces the molecules of the medium to vibrate in a similar fashion and thereby transmit energy in the form of an acoustic wave. Finally, when such an acoustic wave impinges on a material body, it forces the latter into a similar acoustic vibration. In the case of the human ear it produces the sensation of sound. Thus, vibrations in solid and fluid bodies are essential to production, transmission, and reception of sound. *See* MECHANICAL VIBRATION.

One degree of freedom. Systems with one degree of freedom are those for which one space coordinate alone is sufficient to specify the system's displacement from its normal configuration. It is the simplest yet the most fundamental type of vibration system. An idealized example known as a simple oscillator consists of a point mass m fastened to one end of a massless spring and constrained to move back and forth in a line about its undisturbed position (**Fig. 1**). Although no actual acoustic vibrator is identical with this idealized example, the actual behavior of many vibrating systems when vibrating at low frequencies is similar and may be specified by giving values of a single space coordinate. They include loudspeaker cones, telephone diaphragms, microphone diaphragms, and drum membranes. *See* DEGREE OF FREEDOM (MECHANICS).

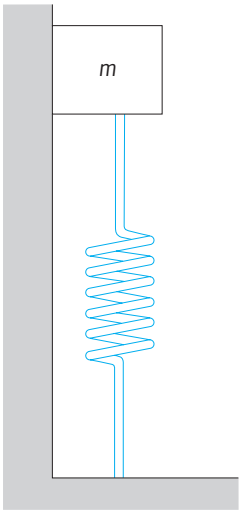


Fig. 1. Simple oscillator.

When the restoring force of the spring of a simple oscillator on its mass m is directly proportional to the displacement x of the latter from its normal position, the system vibrates in a sinusoidal manner called simple harmonic motion. This motion is identical with the projection of uniform circular motion on a diameter of a circle and is represented by the equation $x = A \sin 2\pi ft$. The frequency of vibration f is given by Eq. (1), where s is the constant of proportionality between force and stretch or compression of the movable end of the spring. The constant A represents the amplitude of the vibration, that is, the maximum displacement of mass on either side of its rest position. The magnitude of A is determined by the manner in which the motion is initially started. Note that the frequency of vibration of a simple oscillator is independent of the amplitude of its vibration. *See* HARMONIC MOTION.

$$f = \left(\frac{1}{2\pi}\right)\sqrt{s/m} \tag{1}$$

The variation in velocity v of the point mass of a simple oscillator is given by Eq. (2) and acceleration a is given by Eq. (3). Note that for a given displacement amplitude A, the velocity amplitude $2\pi fA$ is directly proportional to the frequency, and the acceleration amplitude $4\pi^2 f^2 A$ is directly proportional to the square of the frequency. Consequently, large velocity and acceleration amplitudes of a simple oscillator are more readily obtained at high than at low frequencies.

$$v = \frac{dx}{dt} = 2\pi fA \cos 2\pi ft \tag{2}$$

$$a = \frac{d^2x}{dt^2} = -4\pi^2 f^2 A \sin 2\pi ft \tag{3}$$

Energy considerations. The total mechanical energy of a simple oscillator equals the sum of the kinetic energy associated with the moving mass and the

potential energy stored in the distorted spring. When no frictional forces are present, these two types of energy are continuously being converted from one to the other and back again as the vibration cycle progresses. Their sum remains constant at a value given by Eq. (4).

$$E = \frac{sA^2}{2} = 2m(\pi fA)^2 \tag{4}$$

When frictional or other dissipative forces are present, the simple oscillator gradually changes its mechanical energy of vibration into heat energy with an attendant reduction in its amplitude of vibration. No one simple law is capable of describing the variation of frictional forces as a material body vibrates. However, in many important cases the frictional force is opposite to the direction of motion and proportional to the velocity of the vibrating mass. For this type of damping force, the amplitude of vibration of the simple oscillator decreases exponentially with time in accordance with the equation $A = A_0e^{-\alpha t}$, where α is a constant directly proportional to the frictional force and inversely proportional to the mass of the oscillator (**Fig. 2**). The natural frequency of free oscillation of a damped oscillator is slightly less than that of the same oscillator system without damping. The amount by which the frequency is lowered increases with increased damping. *See* DAMPING.

External driving force. A simple oscillator, or some equivalent system, is often maintained in a condition of steady vibration by the application of a periodic external driving force. After a sufficient time has elapsed, any natural free-oscillation frequency of the oscillator dies out and it vibrates solely at the frequency of the impressed driving force. When the applied driving force is sinusoidal, as represented, for example, by the function $F(\cos 2\pi ft)$, the forced vibration ultimately reaches a steady-state amplitude of $A = F/(2\pi fZ)$, where Eq. (5) applies. Z is known

$$Z = \left[R^2 + \left(2\pi fm - \frac{s}{2\pi f}\right)^2\right]^{1/2} \tag{5}$$

as the mechanical impedance of the oscillator. From this equation it is apparent that the driven oscillator will have a maximum amplitude when the forcing frequency is near $[1/(2\pi)]\sqrt{s/m}$, the free oscillation

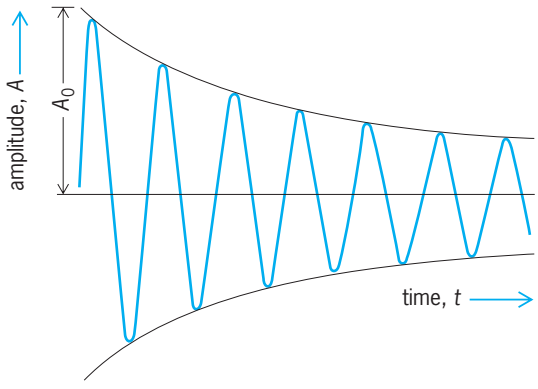


Fig. 2. Damped vibration of simple oscillator.

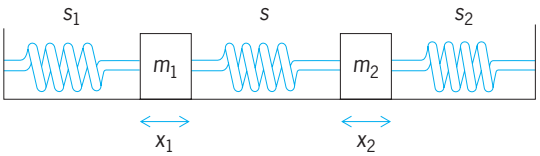


Fig. 3. Simple oscillator with two degrees of freedom.

frequency of the oscillator. *See* FORCED OSCILLATION; OSCILLATOR.

Nonlinear systems. When the amplitude of vibration of any real oscillator system becomes large, the elastic restoring force no longer is proportional to the displacement x. The term s relating force and stretch in the spring is not independent of x and may either increase or decrease as x increases. In either case, the motion no longer is sinusoidal and the frequency is not independent of the displacement amplitude A. The frequency becomes higher when s increases with increasing x and lower for a decreasing s as x increases. The latter characterizes the large-amplitude oscillations of a pendulum bob. *See* PENDULUM.

Two degrees of freedom. When two simple vibrating systems are interconnected by a flexible connection, the combined system has two degrees of freedom. Consider the simple oscillator of mass m_1 and spring s_1 connected to a second oscillator of mass m_2 and spring s_2 by means of the spring s (**Fig. 3**). The motion of m_1 is completely described by the displacement x_1 and that of m_2 by x_2. The system is said to have two degrees of freedom since two independent space coordinates are required to specify the motion.

A system having two degrees of freedom has two normal modes of vibration of respective frequencies f' and f''. Both of these frequencies differ from the respective natural frequencies f_1 and f_2 of the individual uncoupled oscillators. The larger the force constant s of the coupling spring, the larger becomes the frequency difference $(f' - f'')$ relative to $(f_1 - f_2)$; that is, f' increases relative to f_1, and f'' decreases relative to f_2.

Transverse vibrations are defined as those which occur when the vibrations of the medium in question are perpendicular to the direction of propagation of the exciting wave; longitudinal vibrations occur when the vibrations of the medium are lengthwise, along the direction of propagation of the wave. An example of a vibrating system with two degrees of freedom is given by the transverse vibrations of two masses m_1 and m_2 fastened at intermediate points on a tightly stretched wire rigidly supported at each end. Here, the higher frequency mode f' corresponds to a method of vibration in which the individual motions of the two masses are oppositely directed at all times. The lower frequency mode f'' corresponds to one in which the two masses move together in phase, that is, in the same transverse direction at all times.

Several degrees of freedom. A vibrating system is said to have several degrees of freedom if many space

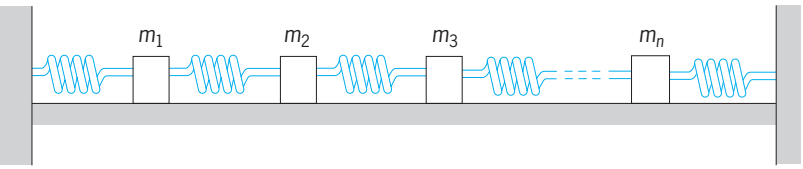


Fig. 4. Vibrating system with several degrees of freedom.

coordinates are required to describe its motion. One example is n masses $m_1, m_2, \ldots, m_n$ constrained to move in a line and interconnected by $(n - 1)$ coupling springs with additional terminal springs leading from m_1 and m_n to rigid supports (**Fig. 4**). This system has n normal modes of vibration, each of a distinct frequency.

Vibrations of elastic bodies. The primary sources of energy for producing sound waves in fluids are vibrations of elastic bodies at frequencies ranging from 20 to 20,000 Hz. In turn, the ultimate detection of sound waves by such devices as the human ear or a microphone requires an elastic body being forced into vibration by the impinging sound waves.

Vibrations of elastic bodies are not usually a simple harmonic motion of just one frequency but instead are of a complex nature, having many natural frequencies and modes of vibration. This complexity arises from numerous factors. Not all the parts of a solid body move together in phase with each other. For instance, a loudspeaker diaphragm has its mass spread over a considerable surface area whose individual sectors may vibrate with different phases and amplitudes with respect to adjacent sectors. Consequently, it often is necessary to give the displacement at each point on the diaphragm as a function of time in order to describe the motion adequately. Most solid bodies are capable of displacements in any direction in space, so that three independent space coordinates are required to specify their vibration. The vibration of a solid body is influenced by interaction with its surrounding medium. The natural frequencies and modes of vibration of a solid body depend upon its shape and dimensions.

It is sometimes possible to confine the mode of vibration to one type having a limited number of natural frequencies. The vibrations of a stretched wire are predominately transverse, although it also is possible to excite weak longitudinal vibrations along the wire. As the diameter of a wire increases relative to its length, the importance of the longitudinal mode of vibration increases until in a thin rod both transverse and longitudinal modes are readily excited. As the diameter of a rod is further increased relative to its length, it ultimately may be regarded as a thin plate capable of vibrating in a thickness mode along its axis or in transverse modes along its surface. In each of these examples it is possible to choose a method of excitation which will encourage one mode of vibration and discourage others.

In any consideration of the vibration of solid bodies, characteristics of major acoustical significance include (1) the natural free vibration frequencies of the body, (2) the segmental vibration pattern for each mode of vibration, and (3) the efficiency with which the vibrations are coupled to the surrounding medium.

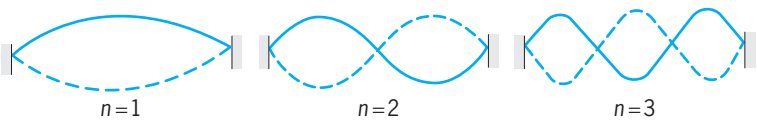


Fig. 5. First, second, and third normal modes of a vibrating string.

Vibrations of strings. The transverse vibrations of thin strings result from tension forces in the string tending to restore any displaced portion of the string to its equilibrium position. The natural frequencies f of a thin string length l rigidly supported at its ends are given by the equation $f = nc/(2l)$, where n may have any integral value 1, 2, 3, ... and c is the velocity with which transverse waves are propagated along the string. In turn, $c = \sqrt{T/m}$, where T is the tension to which the string is stretched and m is the mass per unit length of string. The fundamental, or lowest frequency mode of vibration, of a string is $c/(2l)$. In addition, the string is capable of vibrating at harmonic overtone frequencies of $2\times$, $3\times$, $4\times$, ... this frequency (**Fig. 5**). The simple integral number relationship between the various frequencies of vibration of a string leads to a pleasant harmonic tonal structure which accounts for the use of vibrating strings as the primary source of sound for such musical instruments as the violin, piano, and harp. *See* MUSICAL INSTRUMENTS.

The relative amplitudes of the various harmonic modes of vibration of a string depend upon the particular manner in which the string is initially excited. For instance, if a string is plucked at its center, the even integral (2, 4, 6, ...) harmonic frequencies will be weak compared to the odd integral frequencies. As a string freely vibrates, nodal positions of no displacement are spaced at intervals of l/n along its length. For a string plucked at its center, the even harmonic frequencies are weak since all have a nodal position at the midpoint of the string.

Vibrations of membranes. The transverse vibrations of stretched membranes result from tension forces in the membrane tending to restore any displaced portion of the membrane to its equilibrium position. The theory of vibrations of stretched membranes of either square or rectangular shape is primarily of mathematical interest. However, the vibrations of a circular membrane have certain physical characteristics which find practical applications. These applications include parchment membranes used for drumheads and the stretched thin steel or aluminum diaphragms of condenser microphones.

When a circular membrane is rigidly clamped along its outer radius, it will vibrate at a fundamental frequency given by Eq. (6), where T is the tension to

$$f = \frac{0.384}{a}\sqrt{\frac{Tt}{\rho}} \tag{6}$$

which the membrane is stretched, t its thickness, ρ its surface density, and a its radius. When vibrating

in this mode, the entire surface of the membrane moves in phase with a maximum amplitude at its center. The other possible modes of vibration are not integral multiples of f, but have frequency ratios of 2.295, 3.60, 4.90, and so forth, to the fundamental. When vibrating in these latter modes, one or more nodal circles are present in which the membrane is oppositely displaced on adjacent sides of the circles. The average displacement of the membrane vibrating in this manner is small, because of cancellations between the oppositely phased circular sectors. This condition leads to low acoustical efficiency and limited practical significance for either the production or reception of sound. For example, the response of a condenser microphone at first gradually falls off and then becomes very irregular at frequencies well above the fundamental of its diaphragm.

Vibrations of plates and diaphragms. The transverse vibrations of thin plates or diaphragms result from elastic restoring forces produced when their surfaces are deformed. The natural modes of vibration of such plates are involved functions of boundary shape and dimensions, whether the boundaries are free to vibrate or are rigidly clamped, and the elastic properties of its material. When a thin circular plate is rigidly clamped at its rim, its fundamental frequency is given by Eq. (7), where t is the thickness of the plate, a is its radius, ρ is its density, Y is Young's modulus, and σ is Poisson's ratio for the material of the plate. When vibrating in this mode, the entire surface of the plate moves in phase with a maximum amplitude at its center. The overtone modes of vibration have frequency ratios of 3.88, 8.70, and so on, to the fundamental. When the plate is vibrating in these latter modes, one or more nodal circles are present in which the plate is oppositely displaced on adjacent sides of the circles. The average displacement of the plate's surface under these conditions is small because of cancellations between the oppositely phased circular sectors. As in the case of circular membranes, this condition leads to low acoustical efficiency.

$$f = 0.47 \frac{t}{a^2} \sqrt{\frac{Y}{\rho(1 - \sigma^2)}} \tag{7}$$

The theory of transverse vibration of thin plates finds application in the design of such devices as telephone receiver diaphragms, horn-type loudspeaker diaphragms, underwater sound projectors used in sonar systems, and in understanding the vibrations of wall panels, floors, automobile body panels, and hull plating of ships.

The fundamental mode of vibration of a circular plate free at its rim has a frequency some 13% lower than when rigidly clamped. The relatively involved problem of transverse vibrations of a thin flat plate becomes still further complicated when the plate has a simple curved surface, and a theoretical solution is impossible when the plate is curved into the shape of a bell.

Vibrations of rods. Rod vibrations primarily consist of two simple types: (1) longitudinal vibrations along the long axis of the rod, and (2) transverse vibrations at right angles to this axis.

Longitudinal vibrations. When any plane cross section in a long thin rod is displaced longitudinally relative to adjacent planes, elastic restoring forces caused by either compression or tension in the rod tend to restore the plane to its normal position. These forces result in the propagation of longitudinal waves along the axis of the rod. The interference of two such waves traveling in opposite directions sets up a pattern of standing waves in the rod having certain discrete frequencies. The magnitude of these natural frequencies depends upon the length and material of the rod and upon the particular constraints existing at the two ends of the rod.

If the rod is either free at both ends or rigidly clamped at both ends, its fundamental frequency is given by $f = c/2l$, where l is its length, and c is the velocity with which longitudinal waves are propagated in the rod. When vibrating in its fundamental longitudinal mode, the length of the rod is one-half the wavelength of the waves being propagated along the axis of the rod. Overtone frequencies are given by integral multiples ($n = 2, 3, 4, \ldots$) of the fundamental frequency. In the case of these latter modes of vibration, nodal positions having no longitudinal displacements are spaced at intervals of $l/(2n)$ along the length of the rod.

When the rod is rigidly clamped at one end and free at the other, its fundamental frequency is given by $f = c/(4l)$. Its overtone frequencies are the odd integral ($n = 3, 5, 7, \ldots$) multiples of this frequency. Numerous types of constraints may be placed on the ends of the rod, for example, stiffness of elastic springs, inertia of masses, and so forth. When a rod free at one end is mass-loaded with a mass m at the opposite end, its fundamental frequency gradually decreases from $c/(2l)$ to $c/(4l)$ as the magnitude of the mass increases from zero to infinity.

Whenever a given rod is vibrating longitudinally in one of its natural modes, it may be supported or clamped at a nodal position without interfering with this particular mode of vibration. Since only a few modes of vibration, or in some cases only one, will have a nodal position at a given location, a judicious choice of support position may reduce or eliminate unwanted modes of vibration.

An important practical application for longitudinal vibrators is their use in sonar transducers, where both the magnetostrictive vibrations of nickel tubes and the piezoelectric vibrations of crystals are utilized.

Transverse vibrations. A long rod is capable of vibrating transversely as well as longitudinally, and it is often difficult to produce one motion without exciting the other. When a circular rod is rigidly clamped at one end and free at the other (**Fig. 6**) its fundamental frequency is $f = 0.28ac/l^2$ where c is the velocity of longitudinal waves in the material of the rod, a is its radius, and l its length. The overtone modes of vibration have frequency ratios of 6.27, 17.5, 34.4, and so on, to that of the fundamental.

If the rod were to have a rectangular instead of

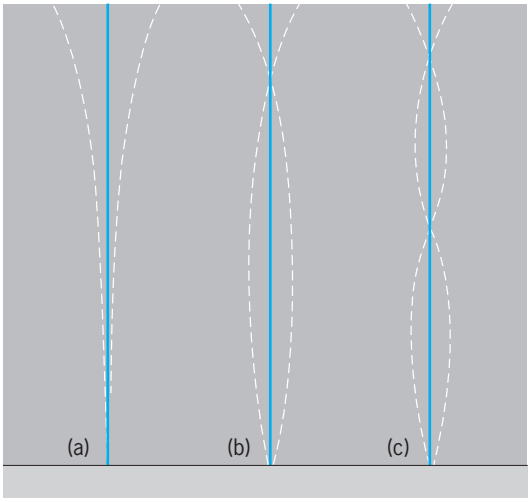


Fig. 6. Transverse vibrations of a long circular rod. (*a*) Fundamental mode. (*b*) First-overtone mode. (*c*) Second-overtone mode.

a circular cross section, its fundamental frequency would be $f = 0.16tc/l^2$, where t is the lateral thickness of the rod in the direction of transverse vibration. Thin rods or reeds of rectangular cross section clamped at one end have numerous applications, including use as a vibrating mouthpiece in the woodwind family of musical instruments and use as resonance vibrators in vibration frequency indicators. The latter type of instrument contains a large number of thin reeds, each having a different resonance frequency. When the base of the indicator is placed in contact with a vibrating object, the free end of that reed having the same natural frequency will be set into vigorous vibration, and its visual motion can be used to indicate on an appropriate scale the numerical value of the vibration frequency.

Another important type of transverse vibration is that of a rod free at both ends. The fundamental frequency of such a rod of rectangular cross section is $f = 0.33tc/l^2$, and there are overtones in ratio to this frequency of 2.75, 5.4, 8.9, and so on. The bars of a xylophone and similar musical instruments vibrate in this manner.

Vibration measuring equipment. Vibrations of solid bodies may result in the generation and transmission of unwanted noise. To determine the source of such noise and to devise methods for its elimination, it frequently becomes necessary to measure these vibrations.

When the vibration of a point in a material body is to be measured, a device must be provided with a sensing element which indicates either the displacements, velocities, or accelerations of the point. Such a device is usually either some mechanical system which indicates the characteristics of the vibration by means of a mechanical pointer, or a pickup capable of converting mechanical energy to electrical or some other form of energy. In conjunction with associated equipment, these pickups may be used to measure solely vibration amplitudes or to give a detailed picture of the entire vibration pattern. A vibrometer, or vibration meter, is a device indicating solely amplitude of vibration, while a vibrograph provides a complete oscillographic record of the vibration.

Vibration meter. A typical electrical vibration meter consists of an electromechanical pickup, adjustable attenuator, amplifier, integrating network, and a direct reading meter capable of being calibrated to read displacement, velocity, or acceleration amplitudes. Connections are also provided for oscillographic presentation, for a pair of headphones for listening to the vibration being measured, for connection to a vibration analyzer, or for connection to an electronic frequency counter.

Mechanical and mechanooptical vibrometers and vibrographs employ gear trains or mechanical or optical lever arms to magnify the vibratory motions before they are indicated or recorded. These devices are usually held by hand, with a probe contacting the vibrating surface (**Fig. 7**). Mechanical vibrometers are normally used at frequencies up to 500 Hz.

Mechanical vibrographs are instruments containing a moving paper or film on which a scribing device records the amplitude of the motion being measured. In one type, shown in **Fig. 8**, the case and recording paper themselves move in proportion to the motion being measured and, for sufficiently high frequencies, a mass supported on a soft spring with an attached scribing stylus serves as a stationary reference. In a second type, the case and recording paper are rigidly fixed, and the vibrations are transmitted

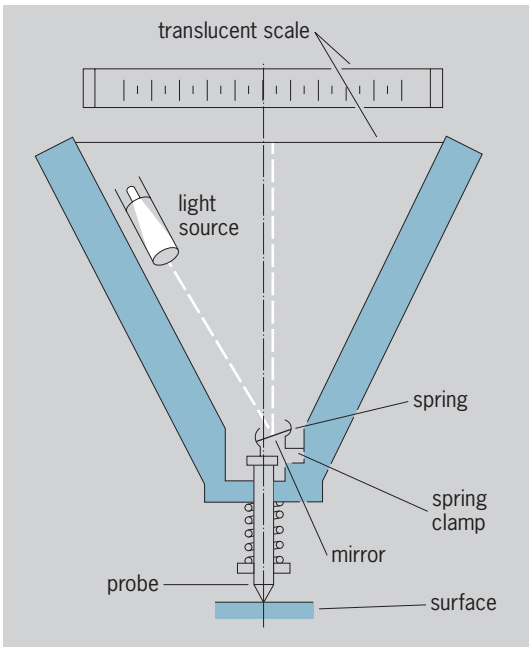


Fig. 7. Mechanooptical vibrometer. The motion given to the probe by the vibrating surface is used to rock a mirror and thereby actuate an optical lever arm. A light beam reflected from the mirror and focused onto the scale provides an indication of the vibration amplitude. (*General Electric Co.*)

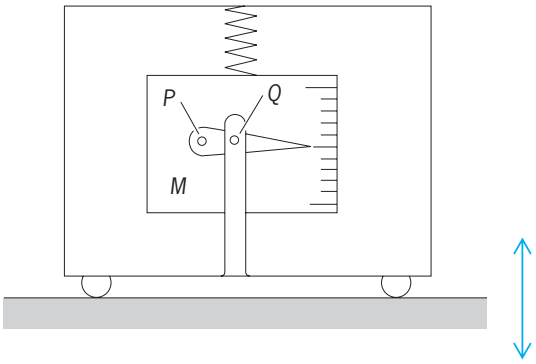


Fig. 8. Mechanical vibrograph. The scribing stylus pivots on point *P*, which is attached to the mass *M*. The case is connected to the stylus at point *Q*. (*After C. M. Harris, ed., Handbook of Noise Control, McGraw-Hill, 1957*)

to the stylus by means of a probe in contact with the vibrating body.

Resonant vibrators. The simple measurement of frequency of vibrating mechanical systems is often made with an instrument containing a series of resonant mechanical vibrators. These sensing elements are a series of cantilever-mounted reeds weighted at their free ends; their natural transverse frequencies are selected to cover a frequency spectrum of interest. Such instruments usually have a relatively small frequency range. By use of a series of these instruments, each covering a different frequency range, an entire range from about 20 to 500 Hz may be covered.

Vibration pickups. These are electromechanical transducers capable of converting mechanical vibrations into electrical voltages. Depending upon their sensing element and output characteristics, such pickups are referred to as accelerometers, velocity pickups, or displacement pickups. *See* ACCELEROMETER; SOUND; VIBRATION DAMPING; VIBRATION ISOLATION; VIBRATION PICKUP; WAVE MOTION.

Lawrence E. Kinsler

Bibliography. J. P. Den Hartog, *Mechanical Vibrations*, 4th ed., 1956, reprint 1985; A. D. Dimarogonas, *Vibrations for Engineers*, 2d ed., 1995; M. Geradin and D. Rixen, *Mechanical Vibrations*, 2d ed., 1997; S. G. Kelly, *Fundamentals of Mechanical Vibrations*, 2d ed., 2000; B. Pippard, *The Physics of Vibration*, 3d ed., 1989; W. Thomson, *Theory of Vibration with Applications*, 5th ed., 1997.

Vibration damping

The processes and techniques used for converting the mechanical vibrational energy of solids into heat energy. While vibration damping is helpful under conditions of resonance, it may be detrimental in many instances to a system at frequencies above the resonant point. This is due to the fact that the relative motion between the base of the vibration isolator and the mounted body tends to become smaller as the isolator becomes more efficient at the higher frequencies. With damping present, the force transmitted by the elastic element is unable to overcome the damping force; this leads to a resulting increase in transmissibility. *See* DAMPING; VIBRATION; VIBRATION ISOLATION.

All metal springs which include structural members such as brackets and shelves have some damping. However, such damping is insufficient for vibration isolators and must be augmented by special damping devices.

Viscous damping. Several different types of damping devices have been developed and used successfully. Probably the most familiar is that used on automobiles, which, although known as a shock absorber, is in reality a damper, and functions as a limiter to the spring system of spring constant k. The system is shown in **Fig. 1**. A piston p is attached to the body m and is arranged to move vertically through the liquid in a cylinder c which is secured to the support s. As the piston moves, the force required to cause the liquid to flow from one side of the piston to the other is approximately proportional to the velocity of the piston in the cylinder. This type of damping is known as viscous damping. The damping force is controlled by the viscosity of the liquid and by the size of the orifice in the piston. There are several disadvantages to this type of damping; for example, it is unidirectional; it is affected by temperature changes; and because the liquid is passed from one side of the piston to the other side through an opening, it is time-conscious. The opening, whether it is an orifice or the clearance between piston and cylinder, can pass only so much liquid in a given length of time. If the body to which the piston is attached is caused to displace faster than the liquid can transfer, a bottoming effect occurs. This effect is experienced by riders in automobiles when a hole in the street is hit at too fast a rate; the springs may appear to bottom out, but actually it is the shock absorber or damper. *See* SHOCK ABSORBER.

Some of the disadvantages of viscous damping may be overcome by using air instead of liquid as the damping medium. Air, being compressible, will add to the effective spring force with large displacements. If the air is housed within a flexible bellows, damping will be attainable horizontally as well as vertically. Such a system is illustrated in **Fig. 2**.

This type of damping has proved very effective

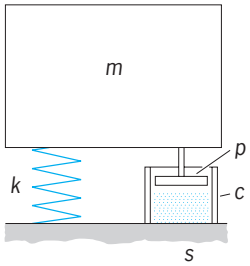


Fig. 1. Automobile shock absorber.

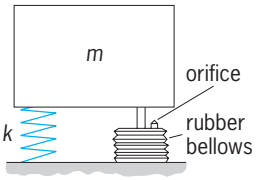


Fig. 2. System employing viscous damping with air.

in vibration isolators. The primary disadvantage occurs under conditions of high and low temperature with the change in elasticity of the rubber bellows.

Friction damping. Damping forces may be generated by causing one dry member to slide on another. This is known as dry friction or coulomb damping. A damper of this type is shown in **Fig. 3**. A pin p inserted in cylinder c and attached to body m is arranged to slide between two vertical spring members which are attached to the support s. The force exerted by the damper in opposition to the motion of the body m is the product of the normal force and the coefficient of friction between the pin and vertical leaf springs. The damping force is usually constant; however, if the pin is tapered, a variable force may be obtained. Friction damping is used in several commercially available isolators because it provides a simple means to control the damping forces. If necessary, independent damping systems may be provided within the same unit for vertical and horizontal motions. Some frictional dampers are effective in both vertical and horizontal directions. One such system (**Fig. 4**) comprises a load-carrying concave-convex spring made of a metal screen consisting of two parts attached by an eyelet. A damping-coil spring encircles the two attached springs. Any motion vertically

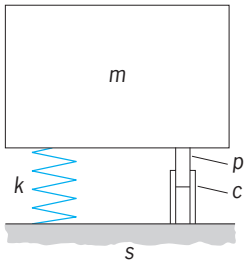


Fig. 3. System employing Coulomb damping.

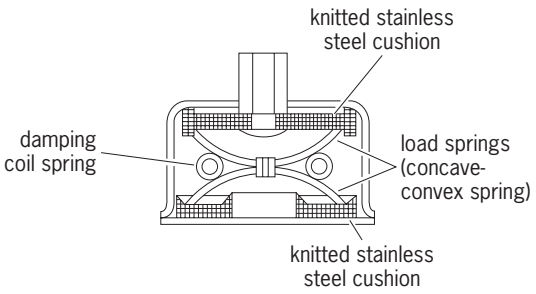


Fig. 4. Frictional damper.

or horizontally will cause the load springs to deflect vertically or tilt horizontally. Any change in position of the load springs results in the damping spring's being forced out over the surface of the concave springs; this creates a frictional force. Since the surface contact of the damping spring increases with displacement, the damping force is approximately linear.

Inherent damping. There are many applications where vibration dampers of an external type such as those discussed cannot be used because of space limitations, economic considerations, or the fact that the system needs very little damping. These applications make use of the inherent damping, or internal hysteresis, of such materials as rubber, felt, and cork. Vibration isolation with inherent damping is most commonly used in applications with constant motor speeds, such as air compressors, generators, and grinders.

Magnetic damping. This type of damping is attainable as a result of the electric current induced in a conductor moving through a magnetic field. The damping force can be made proportional to the velocity of the conductor moving through the field. Magnetic damping has not been used successfully in vibration isolators because its effectiveness is limited to a single direction. *See* ELECTROMAGNETIC INDUCTION.

Numerical values. Considerable emphasis is placed upon the numerical value for the damping force. Such a value is needed to predict the behavior of a damped system; however, difficulty is often encountered in such an analysis. The difficulty is that viscous damping, although susceptible to mathematical analysis and the establishment of a numerical value for the damping force, is seldom encountered in pure form in actual practice. The types of damping actually used are not well adapted to mathematical analysis. The damping value may be determined by the logarithmic decrement method, but care should be exercised when these values are used in equations of pure viscous damping. This is especially true of the two most common forms of damping—friction and inherent damping. With both these types the effect of damping will vary with the amplitude of vibration. Since the damping effect is different, the natural frequency of the system will change with amplitude of vibration. For double amplitudes of vibration, such as 0.060 in. (0.15 cm), the magnification at resonance may be one value and the resonance frequency may be so many cycles per second, while with a double amplitude of 0.020 in. (0.05 cm), the magnification may be different and the resonance period may be higher. This difference must be recognized when applying formulas derived from the analysis of pure viscous damping. *See* SPRING (MACHINES); VIBRATION MACHINE.

K. W. Johnson

Bibliography. L. L. Beranek and I. L. Ver, *Noise and Vibration Control Engineering: Principles and Applications*, 1992; J. P. Den Hartog (ed.), *Mechanical Vibrations*, 4th ed., 1956, reprint 1985; C. M. Harris, *Shock and Vibration Handbook*, 4th ed., 1995;

A. D. Nashif, D. I. Jones, and J. P. Henderson, *Vibration Damping*, 1985; S. S. Rao, *Mechanical Vibrations*, 2d ed., 1990.

Vibration isolation

The isolation, in structures, of those vibrations or motions that are classified as mechanical vibration. Vibration isolation involves the control of the supporting structure, the placement and arrangement of isolators, and control of the internal construction of the equipment to be protected.

The simplest kind of mechanical vibration has the waveform of sinusoidal motion (**Fig. 1**). Vibrations in structures, although generally more complex in waveform, exist wherever movement takes place. Such movement may be caused, for example, by the engine in an automobile, by engines or wind buffeting in aircraft, or by a punch press in a building. Delicate electronic equipment and precision instruments must normally be isolated from these motions if accurate measurements are to be obtained. *See* MECHANICAL VIBRATION; VIBRATION.

Vibration, in most cases, may be effectively isolated by placing a resilient medium, or vibration isolator, between the source of vibration and its surrounding area to reduce the magnitude of the force transmitted from a structure to its support or, alternatively, to reduce the magnitude of motion transmitted from a vibrating support to the structure. Isolating vibration at its source is commonly termed active or source isolation; isolating an instrument from its surroundings is known as passive isolation. In either case the vibration isolator is designed according to the same principles. Vibration isolators are available commercially, with published data covering their characteristics. These data include minimum and maximum rated load, natural frequency at rated load, transmissibility, damping characteristics, ultimate strength, sway space limits, and so on.

Natural frequency; resonance. The prime concern in the field of vibration isolation is the proper use of isolators under various load configurations with respect to their loading, the desired natural frequency, the position and location of the isolators, and the relationship to the structural response of the equipment to which they are attached. For the vibration isolator to be effective, the natural frequency should be approximately 0.4 times the frequency of the interfering source. The natural frequency is the frequency at which a freely vibrating mass system oscillates once it has been deflected. In the case where vibrations occur over a wide frequency range, such as in aircraft (5–2000+ Hz), the natural frequency of the isolator is established with respect to the cruising speed. This means that when the lower frequencies are traversed, such as during aircraft takeoff, the mounted equipment will pass through a condition known as resonance. Resonance is said to exist when the natural frequency of a spring (in this case an isolator) coincides with the frequency of the excitation forces. Resonance causes magnification of the input vibration and may be harmful to the equipment if not controlled within reasonable limits. To control the vibration magnification at resonance, the resilient element within the isolator must be damped. With suitable damping, the magnification factor may be held to 3 or less. *See* DAMPING; RESONANCE (ACOUSTICS AND MECHANICS); VIBRATION DAMPING.

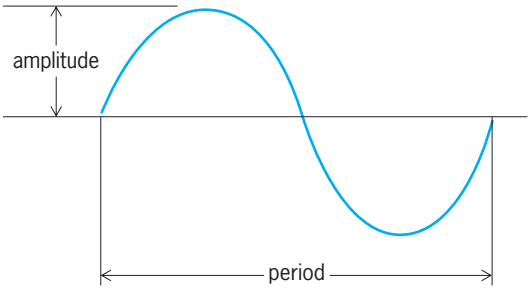


Fig. 1. Sinusoidal motion.

The vibration isolator should be considered as only one part of the isolating system, the other parts being the supporting structure that lies below the isolator, and the internal structure of the equipment that is above the isolator. When isolators are selected for use where the period of resonance is critical, it should not be forgotten that the flexibility of the supporting structure and the flexibility of the isolators are in series, so that the resultant resonant frequency of the loaded system will therefore be inversely proportional to the square root of the sum of these two flexibilities. The additional flexibility of the structure will lower the natural frequency of the system and will also result in increased displacement during resonance, caused by the presence of the undamped structure. The flexibility of the supporting structure, including the structural linkages leading to such structures, should not exceed 25% of the flexibility of the isolator. To neglect the resiliency of the support structures in providing the desired natural frequency of the system, as so many textbooks do with the assumption of rigid support, is impractical.

There are several methods for determining the load at each isolator location; one is shown in **Fig. 2**, where the loads are as given in the following

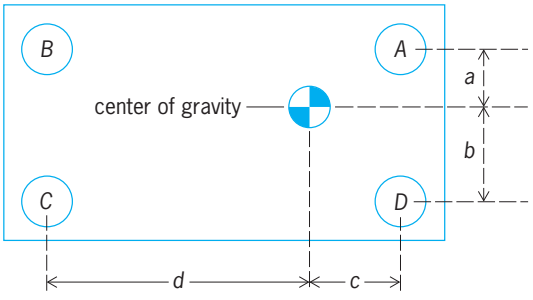


Fig. 2. Determination of isolator locations.

four equations (W is the total weight).

$$A = W\left(\frac{b}{a+b}\right)\left(\frac{d}{c+d}\right)$$
$$B = W\left(\frac{b}{a+b}\right)\left(\frac{c}{c+d}\right)$$
$$C = W\left(\frac{a}{a+b}\right)\left(\frac{c}{c+d}\right)$$
$$D = W\left(\frac{a}{a+b}\right)\left(\frac{d}{c+d}\right)$$

The next step in vibration isolation is the positioning and arrangement of the isolators with regard to the geometry of the equipment.

Location of isolators. The vibration isolators may be positioned and arranged in many different ways, all variations of three basic types, each of which requires a definite amount of space: (1) isolators attached underneath equipment, known as an underneath mounting system; (2) isolators located in the plane of the center of gravity of the equipment, known as a center-of-gravity system; (3) mountings arranged four on each side in the plane of the radius of gyration, known as a double side-mounted system or radius-of-gyration system. *See* CENTER OF GRAVITY; RADIUS OF GYRATION.

Textbooks generally treat problems relating to the mounting of equipment with resilient elements as masses in unlimited space. Under these undefined conditions any of the three systems listed may be used; however, when a space limitation is imposed, each system has a definite limitation and must be used accordingly.

Underneath mounting system. For an underneath mounting system, the most efficient for vibration isolation is one with a low natural frequency in both the vertical and horizontal axes. The most stable system is one with a high natural frequency with respect to the disturbing frequency, but in this case the vibration characteristics will suffer. Stability may be achieved with low-frequency isolators by maintaining adequate spacing of the isolators with respect to the height of the center of gravity from the mounting plane. Tests show that the isolator spacing should not be less than twice the height of the center of gravity from the mounting plane. This condition is illustrated in **Fig. 3** . The use of an underneath mounting system for equipments that exceed the condition shown in Fig. 3 should be augmented by stabilizers to provide the needed stability.

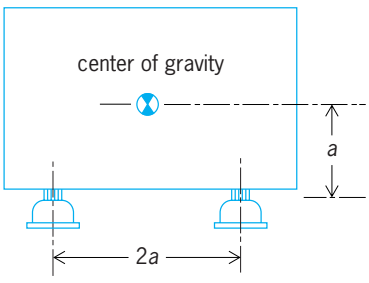


Fig. 3. Underneath mounting system.

Center-of-gravity system. Locating the isolators in the plane of the center of gravity has generally been considered the ideal system because of its ability to decouple the rotational modes of vibration. In actual use, however, such results are seldom achieved because of space limitations. The primary conditions are (1) that the isolators be located in a plane passing through the center of gravity, (2) that the distance between isolators be twice the radius of gyration of the body, and (3) that the horizontal-to-vertical stiffness of the isolators be equal. The first and last of these conditions are easy to satisfy. The second condition is not always possible where a space limitation exists. The limit of the center-of-gravity system, especially in aircraft, is generally arrived at when the height-to-width ratio of the equipment exceeds 2.8.

Radius-of-gyration system. The third system that may be used when the limit of the center-of-gravity system has been reached is a double side-mounted system or radius-of-gyration system as illustrated in **Fig. 4**. Two sets of isolators are arranged on each side. For optimum results the isolators should be located in the plane of the radius of gyration. However, since it is often difficult to determine the exact radius, acceptable results will be obtained in most instances by assuming the body to be of uniform density. Satisfactory results have been obtained with bodies having height-to-width ratios up to 5. The limitation of this system will be reached when structural rigidity of the body is such that excessive bending occurs between the upper and lower isolator locations.

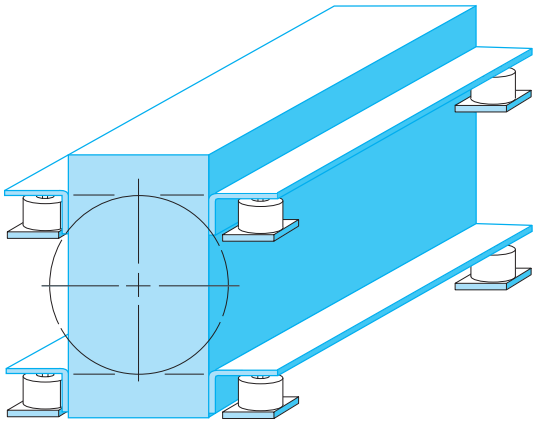


Fig. 4. Radius-of-gyration system.

The next consideration in vibration isolation is the structural rigidity of the body to be isolated. This step is of importance in that the use of incorrect structures, particularly supporting brackets for component parts, can render the other two steps useless. Supporting brackets act as springs under a vibratory condition and become resonant at their natural frequency. Should resonance of the brackets coincide with the isolator resonance, damage may occur. The resonance point for the brackets and internal component should occur when the isolators are approaching their maximum efficiency so that the

input vibration to the internal structure is at a low level. This level is normally reached at four times the natural frequency of the isolator.

The isolation of vibration may be accomplished in most instances by the use of commercial isolators. In some cases, such as in heavy industrial machinery where the supporting structure is concrete, less consideration need be given to the flexibility of the supporting structure. Also, where the body to be isolated is stationary, as in the case of compressors, the height-to-width ratio with respect to the isolator spacing is of little importance. *See* VIBRATION MACHINE.

K. W. Johnson

Vibration machine

A device for subjecting a system to controlled and reproducible mechanical vibration. Vibration machines, commonly called shake tables, are widely used in vibration measurement and analysis. There are three types of vibration machines in general use. These are the mechanical direct-drive type, the mechanical reaction type, and the electrodynamic type. Other types, such as hydraulic excitation devices, resonant systems, piezoelectric vibration generators used for instrument calibration, and machines for testing packages, have only limited or specialized applications.

Direct-drive machine. The mechanical direct-drive vibration machine (sometimes referred to as the brute-force type) is a machine in which the vibration table is forced by a positive linkage to undergo a displacement. The linkage is driven by a direct attachment to eccentrics or crankshafts. The degree of eccentricity or crank radius determines the amplitude of vibration of the table. The motion of the vibration table is approximately simple harmonic. The frequency is varied over a range of 5–60 Hz by a variable-drive unit. Automatic cycling is normally used to cycle the table to maximum frequency and return to the starting frequency. The period of cycling is usually 1–3 min. *See* HARMONIC MOTION.

The vibration table is usually designed, by the use of several cam linkages, to move in both vertical and horizontal directions. A machine of this type is pictured in **Fig. 1**.

Reaction-type machine. A vibration machine in which the forces exciting the vibrations are generated by rotating or reciprocating unbalanced masses is called a reaction-type machine. The system consists of a table which is suspended from a mounting frame attached to the floor. The unbalanced masses which are attached to a shaft are part of the floating table assembly. A separate variable-speed unit, which is attached to the mass-shaft assembly of the table by a flexible coupling, drives the system.

The amplitude of vibration is controlled by the degree of off-center setting of the rotating masses. Frequency is controlled by a rheostat in series with a dc motor. The frequency may be continuously varied over a range of 5–60 Hz. The primary advantage of this type of machine is the constant displacement provided by the rotating masses. Also, since the table is suspended, there is essentially no vibration at the floor attachment points. A reaction-type machine is pictured in **Fig. 2**.

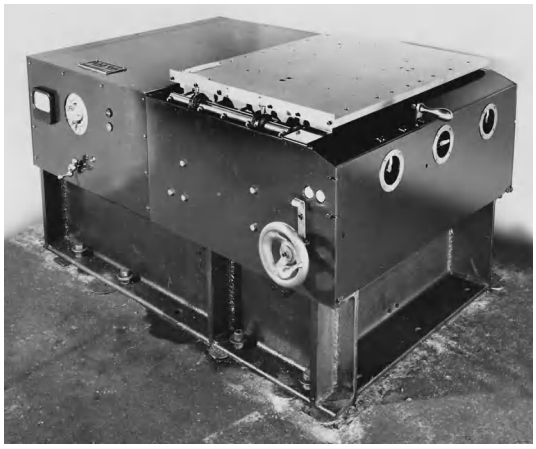


Fig. 1. Direct-drive vibration machine. (*L. A. B. Corp.*)

Electrodynamic machine. The electrodynamic vibration machine derives its vibratory force from the action of a fixed magnetic field upon a coil of wire contained in it and excited by a suitable alternating current. The armature consists of those parts that are integral with the coil and the associated elements that move in the magnetic field. This type works on the same principle as that of an electrodynamic loudspeaker. The force output is proportional to the magnitude of the current. Automatic displacement or acceleration control may be had over a frequency range of 5–2000+ Hz.

The electrodynamic vibration machine provides motion of a sinusoidal waveform and a form commonly termed random vibration.

Fig. 2. Reaction-type vibration machine. (*L. A. B. Corp.*)

Fig. 3. Complete vibration test system showing the electrodynamic shaker in the left foreground. In the rear, from left to right, are the power amplifier, control console, and instrumentation cabinets. (*Calidyne Co.*)

The primary advantage of this type of vibration machine is the wide frequency range it offers. **Figure 3** pictures a machine complete with the control console. *See* MECHANICAL VIBRATION; VIBRATION.

K. W. Johnson

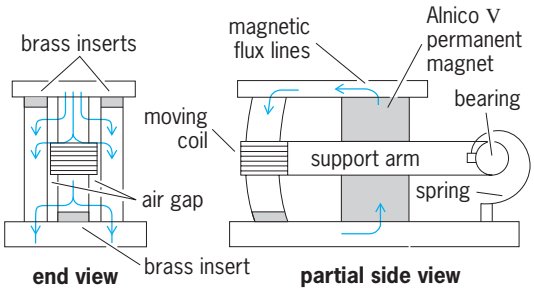
Vibration pickup

An electromechanical transducer capable of converting mechanical vibrations into electrical voltages. Depending upon their sensing element and output characteristics, such pickups are referred to as accelerometers, velocity pickups, or displacement pickups.

Accelerometer. This device consists essentially of a mass which is seismically supported with respect to a surrounding case by means of a spring and guided to prevent motions other than those along the seismic direction of support. In the operating frequency range of the accelerometer, which is below its resonant frequency, the mass experiences essentially the same acceleration as does the case of the accelerometer. The mass exerts a force on the spring's support which is directly proportional to the acceleration being measured. This, in turn, is converted into an electrical voltage by means of stresses produced in a piezoelectric crystal. *See* ACCELEROMETER.

Velocity pickup. This device generates a voltage proportional to the relative velocity between two principal elements of the pickup, the two elements usually being a coil of wire and a source of magnetic field (see **illus.**). In a probe type of velocity pickup, the relative motion is between a coil, which is attached to one end of a probe having its other end in contact with the structure whose vibration is being measured, and a stationary magnetic field produced inside the hand-held case. The magnetic field is essentially stationary in space so that the voltage output of the moving coil is directly proportional to the velocity of the vibrating body.

Displacement pickup. This is a device that generates an output voltage which is directly proportional to the relative displacement between two elements of the instrument. These pickups are similar in construction and behavior to velocity pickups. The only essential difference is the use of a frequency-weighting network, required to make them direct-reading. Instrumentation problems in vibration measurements are generally reduced if measurements of small displacements at high frequencies are made



Velocity pickup. The coil swings on one end of an arm which is supported by bearings at the opposite end. The case follows the motion of the structure to which it is attached. (*MB Manufacturing Co.*)

with accelerometers or velocity pickups, since for a given displacement the acceleration varies as the square of the frequency and the velocity directly as the frequency. *See* VIBRATION. Lawrence E. Kinsler

Vibrotaction

The response of tactile nerve endings to varying forces on the skin and to oscillatory motion of the skin. Grasping, holding, and tactile exploration of an object are part of everyday experience. The performance of all these activities is dependent on the dynamic response of specialized nerve endings located in the skin.

Knowledge of the neural and psychophysical processes involved in vibrotaction is necessary to develop effective tactile communication systems, such as vibrotactile pagers, and multipin vibrotactile stimulators used to produce images on an extended skin surface. Potential users range from business people to the visually and hearing impaired: potential applications range from activities involving visual and aural sensory saturation (such as pilots in high stress situations) to those requiring sensory stimulation (for example, virtual reality).

The success of vibrotactile devices and the successful performance of manual tasks involving grasping and holding small or delicate objects ultimately depend on maintaining adequate tactile function.

Touch receptors. Small electrical impulses, or action potentials, generated by a single nerve ending can be recorded by inserting a microelectrode into the arm of an alert human subject and positioning the electrode tip in a nerve fiber. Studies of the action potentials produced when the skin is locally depressed at a fingertip have identified activity in four networks of distinctive nerve endings lying within 1–2 mm of the skin surface. These nerve endings, which transform skin motion into neural signals, consist of physically and functionally different mechanoreceptors.

Three types of mechanoreceptor in the fingertip have been implicated in vibrotaction. One of these is a population of slowly adapting mechanoreceptors (SA type I, or SAI); their action potentials persist for some time after a transient skin indentation. The two others are populations of rapidly adapting receptors (FAI and FAII types). Type I receptors respond only to skin indentation within a few millimeters of the nerve ending, whereas type II receptors also respond to stimuli at a greater distance from the nerve ending. A fourth mechanoreceptor type appears to respond primarily to skin stretch (denoted SAII).

Mechanoreceptor populations. Exploration of the surface of an object by the fingertips produces spatially and temporally varying tactile stimuli.

For example, to record the responses of single human mechanoreceptors to the Braille representation of letters, a sequence of embossed dots is scanned repeatedly over a (stationary) fingertip in one direction. Between each scan, the sequence of dots is shifted in a direction normal to the scan. The action potentials generated by the skin-stretch receptors and FAII receptors do not resolve the Braille characters. In contrast, the characters are clearly resolved by the SAI receptors, and are partially resolved by the FAI receptors. Thus, surface topography of this form and scale can be detected peripherally by individual nerve endings. The SAI receptor population also establishes the limit of resolution of gross surface features, such as edges and cracks.

When scanned by the fingertip, small-scale surface features or textures (for example, sandpaper or silk) generally require more complex neural representation in order to be detected. In the detection of surface texture, the slowly adapting and the rapidly adapting receptor populations are implicated, either together or separately, depending on the nature of the surface and details of the task. For example, the texture of finely etched surfaces, consisting of dots 1 μm in relief, 50 μm in diameter, and separated by 100 μm, can be detected when stroked across the fingertip. In this case, the texture appears to be detected by neural activity in the FAII receptor population.

Successful lifting and gripping of small or delicate objects between the thumb and fingertips (for example, an eggshell) are also dependent on vibrotaction. This is because precision grip is controlled indirectly by neural activity in the slowly adapting and rapidly

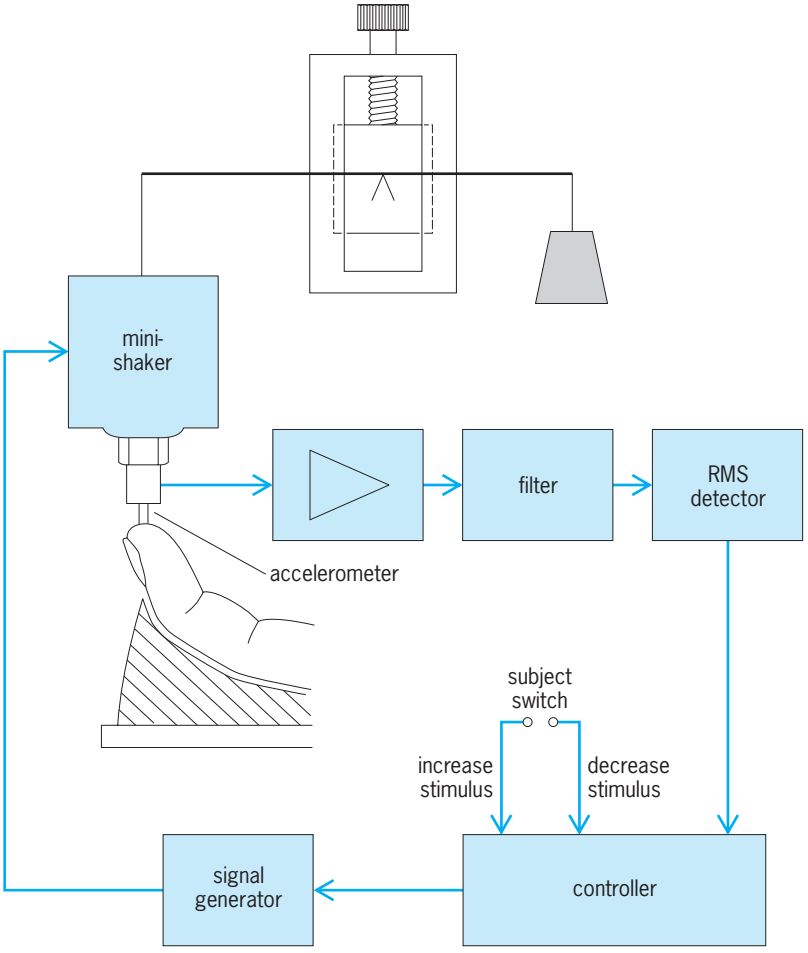


Fig. 1. Block diagram of apparatus for determining vibrotactile perception thresholds at the fingertip.

adapting mechanoreceptor populations, which triggers compensatory muscle action in response to the detection of microscopic slippage of the object.

Thresholds. The existence of four, separate neural channels at stimulation levels close to the threshold of vibrotactile perception has been confirmed by psychophysical experiments. Three of these neural channels, corresponding to the slowly adapting type 1 and both types of rapidly adapting mechanoreceptors, can determine independently the vibrotactile threshold in specific frequency ranges with appropriate stimulation. Therefore it is possible to express changes in mechanoreceptor acuity in terms of changes in vibrotactile perception thresholds at selected frequencies.

The determination of the acuity of the three mechanoreceptor populations requires apparatus such as that shown in **Fig. 1**. The subject is seated comfortably with forearm and hand supported horizontally by a contoured arm rest. A small-diameter, flat-ended, cylindrical probe is lowered onto the fingertip and positioned between the center of the whorl and the nail. The stimulator is mounted on a beam balance (the fulcrum of which may be raised or lowered) so as to maintain the extremely small contact force required between the skin and stimulating probe (typically 0.15 N or 0.5 ounce-force).

Sinusoidal bursts of vibration followed by quiescent periods are then applied to the fingertip, with amplitude initially increasing between successive bursts. Detection of the tone burst by the subject is signaled by pressing a switch, whereupon the stimulus amplitude is recorded by a computer. The vibration intensity is then decreased between bursts until the stimulus can no longer be felt, whereupon the stimulus amplitude is again recorded and the amplitude progression again reversed.

Tactograms. The results of vibrotactile threshold measurements may be conveniently summarized by a tactogram (**Fig. 2**). This graphical representation of the data is equivalent to the audiogram obtained from audiometric measurement of hearing acuity. A tactogram shows the change in threshold sensitivity from the mean value for healthy hands as a function of stimulation frequency. Thus, it can measure the differences in acuity between fingers and between individuals. A tactogram can also measure the losses in acuity resulting from injury, disease, or aging. *See* AUDIOMETRY.

One type of abnormal tactogram is characterized by similar reductions in threshold sensitivity at all frequencies and is exemplified by the data plotted as circles in Fig. 2. This similar loss in acuity occurring in each network of mechanoreceptors and associated nerve fibers is compatible with changes occurring within the whole nerve. A second type of abnormal tactogram involves elevated thresholds at frequencies mediated by only one receptor type (for example, the data plotted as squares in Fig. 2). A mechanoreceptor-specific mechanism is unlikely to occur within a nerve trunk; therefore this frequency-selective loss pattern is suggestive of neural dysfunction occurring at, or close to, the nerve endings. This pattern of abnormal threshold elevation has been observed only in the hands of manual workers who operate vibrating power tools. *See* CUTANEOUS SENSATION; MECHANORECEPTORS.

Anthony J. Brammer

Bibliography. S. Bolanowski et al., Four channels mediate the mechanical aspects of touch, *J. Acous. Soc. Amer.*, 84:1680–1694, 1988; A. J. Brammer and J. E. Piercy, Rationale for measuring vibrotactile perception at the fingertips as proposed for standardization in ISO 13091-1, *Arbetslivsrapport*, 4:125–132, 2000; K. O. Johnson and S. S. Hsiao, Neural mechanisms of tactual form and texture perception, *Annu. Rev. Neurosci.*, 15:227–250, 1992; M. A. Srinivasan, J. M. Whitehouse, and R. H. LaMotte, Tactile detection of slip: Surface microgeometry and peripheral neural codes, *J. Neurophysiol.*, 63:1323–1332, 1990.

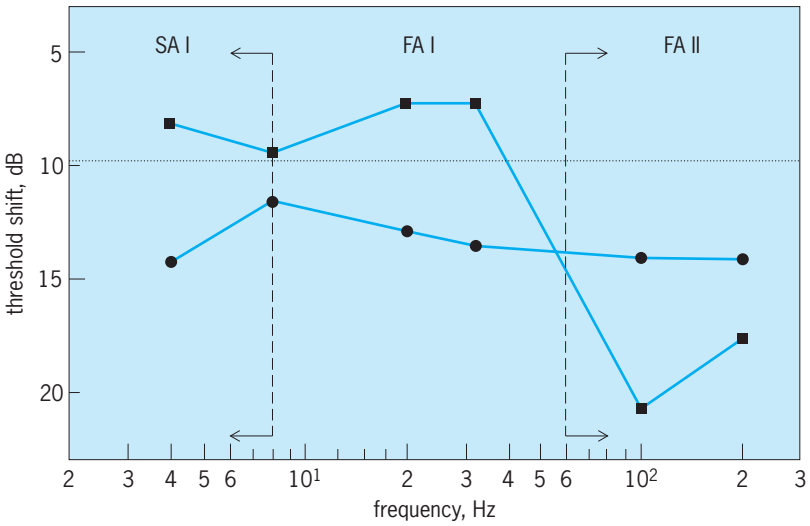


Fig. 2. Tactogram showing the approximate maximum threshold recorded at the fingertips of healthy persons (horizontal line), and examples of two types of threshold elevation (reduction in acuity) at selected frequencies (circles and squares).

Vicuna

A rare animal whose fiber makes the world's most costly and most exquisite cloth, surpassing all others in fineness and beauty. It is found in an almost inaccessible area of the Andes Mountains, at altitudes between 16,000 and 19,000 ft (4876 and 5791 m). The vicuna, one of the wildest of animals, is less than 3 ft (0.9 m) high and weighs 75–100 lb (33.7–45 kg).

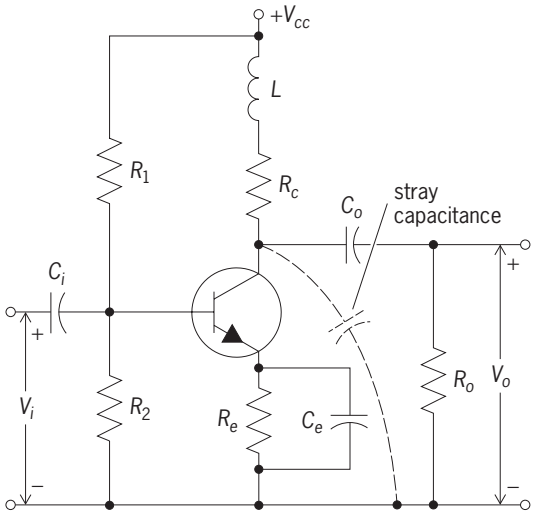
A single animal yields only approximately $^1/_4$ lb (0.11 kg) of hair; thus 40 animals are required to provide enough hair for the average coat. To preserve the species, the Peruvian and Bolivian governments have placed the vicuna under their protection. Attempts to domesticate this animal have not been successful but are still being made in Peru. The fiber of the vicuna is the softest and most delicate of the known animal fibers; yet it is strong for its weight, is resilient, and has a marked degree of elasticity and surface cohesion. It is used in costly suitings and overcoat fabrics. *See* ALPACA; CAMEL'S HAIR; CASHMERE; LLAMA; MOHAIR; NATURAL FIBER; WOOL.

M. David Potter

Video amplifier

A low-pass amplifier having a bandwidth in the range from 2 to 100 MHz. Typical applications are in television receivers, cathode-ray-tube computer terminals, and pulse amplifiers. The function of a video amplifier is to amplify a signal containing high-frequency components without introducing distortion.

In a single stage of an *RC* (resistance-capacitance) coupled amplifier the high-frequency half-power limit is determined essentially by the load resistance, the internal transistor capacitances, and the shunt capacitance in the circuit. To extend the bandwidth of an *RC*-coupled amplifier, it is necessary to overcome the effects of these capacitances. In the past, this was often done by adding an inductance *L*, as shown in the **illustration**, employing shunt peaking. The principal design requirement is to choose a value for the inductance that will extend the frequency response without introducing an undesirable hump in the gain characteristic for frequencies near the upper half-power frequency.



Shunt-compensated amplifier, in which effect of capacitances is overcome by adding inductance *L*.

Modern video amplifiers use specially designed integrated circuits. With one chip and an external resistor to control the voltage gain, it is possible to make a video amplifier with a bandwidth between 50 and 100 MHz having voltage gains ranging from 20 to 500. The use of integrated video amplifiers minimizes cost and space. These amplifiers also eliminate the need for the shunt-peaking inductor shown in the illustration. *See* AMPLIFIER; INTEGRATED CIRCUITS.

Harold F. Klock

Bibliography. D. Christiansen (ed.), *Electronics Engineers' Handbook*, 4th ed., 1996; J. Millman and A. Grabel, *Microelectronics*, 2d ed., 1987; A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, 4th ed., 1997.

Video disk

A medium used to record, distribute, and play video information. The video disk exists in three major forms: the Digital Video Disk (DVD; sometimes also called the Digital Versatile Disk); the Laser Video disk (sometimes called the Laser Vision disk); and the Video CD. There is an audio version of the DVD, called the Digital Audio Disk, that is based on the same digital principles as the DVD. *See* COMPACT DISK.

The advantages of DVD and Laser Video disks over videocassette tapes include very high video quality (because the signal-to-noise ratio is very high); use of a full television frame on playback; multiple audio channels, digital and analog; archival quality, since there is no disk deterioration over time; the possibility of interactivity, whereby cues, branches, and so forth can be programmed in; a freeze-frame option with no loss of quality; the ability to interface and interact with computers using digital cues; the availability of rapid access to any point in the recording; the possibility of auto-stop, whereby a frame is preselected for freezing; frame-by-frame viewing, providing slow motion with no loss of quality; and programmability with certain systems. The Video CD does not possess these advantages, but is a lower-cost option with limited image quality and minimal interactivity.

Common elements. All three types of video disks have basic elements in common, notably a solid-state laser a fraction of a square millimeter in size that emits a few milliwatts of coherent light focused to a micrometer-sized spot (see **illus.**). This laser spot becomes essentially an optical stylus that performs various functions such as tracking, focusing, and reading the total integrated light. *See* LASER.

Tracking and focusing. The laser spot follows a track on the rotating disk that is either the data itself or a groove. It does this by sensing the symmetry of the laser light reflected back to a split segmented detector; if the spot is not symmetric, the position of the head is jogged until it is back to a perfectly registered spot on the detector. The jogging movement is created by a small motor, often a voice coil motor, that is driven by an electronic servo system that "closes the loop" in an iterative manner. The laser spot and the disk are kept in focus through a similar servo-based controller that uses the image of the laser on the detector.

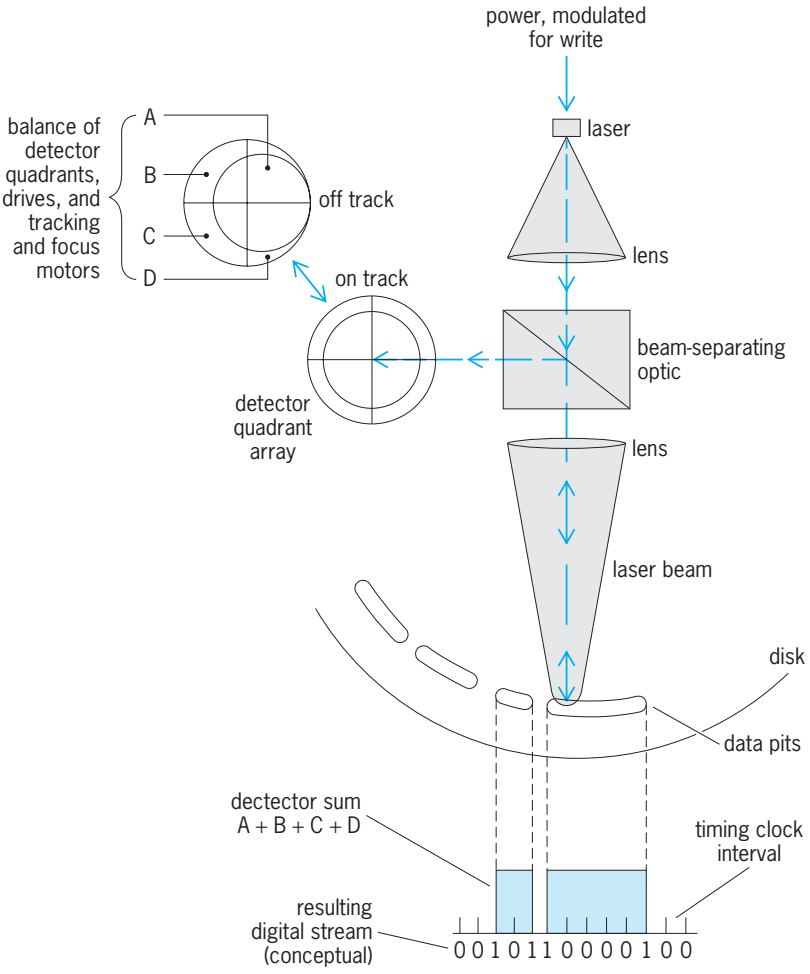
Total integrated light. By adding up the total integrated light on the detector, it is possible to read the total light returned from the disk. While the laser light travels over the disk, differences in the intensity of the light, affected by the disk features that represent the data, are detected. These disk features can be pits in the surface (using the phase contrast of light coming off the disk) or the absorbed light in a dye or other material. This change in intensity is detected and the pulse train decoded by the system to produce a series of data bits that are then reconstructed to form a video image. This principle is used by all the above video disk applications. *See* OPTICAL RECORDING.

TABLE 1. Major changes to CD technology that created DVD

Parameter changed	Result
Code format changed from EFM to EFM+	Increased density × 1.1
Track pitch changed from 1.6 to 0.74 micrometers	Increased density × 2
More efficient error correction (ECC)	Increased density × 1.3
Smaller minimum feature (400 nm)	Increased density × 2

DVD. This is a video version of the original compact disk (CD) and has many similarities. As for CDs, the data that are encoded on the disk, and later decoded, are digital rather than analog. This means that the discrete disk pits are decoded to give a sequence of 1's and 0's, which then can be manipulated using digital techniques. The DVD uses the same conceptual method for tracking and focusing and the same constant linear velocity (CLV) rotation method as CDs. The disk dimensions are the same, with 120-mm (4.72-in.) and 80-mm (3.15-in.) outer-diameter versions. However, the DVD disk itself is only 0.6 mm thick (versus 1.2 mm); two are then laminated together to get 1.2 mm as in CDs. The inner hole diameters of DVDs and CDs are identical at 15 mm (0.6 in.).

Significant changes were made in DVD disks relative to CDs, including increased density and data rate. Through those changes and by the use of MPEG (Motion Picture Enhanced Graphics) compression technology, it was possible to create 4.7-GB disks that can contain several hours of, for example, an NTSC (National Television System Committee) quality movie. The development of data compression technology was critical to the success of DVD; it allows a single 4.7-GB disk to contain up to two versions of a typical full-length movie as well as supplemental information such as parallel sound tracks. This has been pivotal to its commercial success. Newer versions of the DVD disk have higher capacities and also are recordable (**Table 1**). *See* DATA COMMUNICATIONS; DATA COMPRESSION.



DVD operational principles.

Eight to Fourteen Modulation (EFM) code, the way that a collection of 1's and 0's is translated to bytes, contains rules that must be followed to ensure that overlap of disk pits does not occur. The change to EFM+ for DVDs produced a more efficient translation method. The error correction (ECC) method used in DVD is called the 2Dimensional Product Code and is able to reduce errors on the disk from one error in 1000 bits to virtually none. The changes in track pitch and minimum feature size (the smallest feature that can be read) resulted from changing the wavelength of the laser from 780 nm to 650 nm (the shorter the wavelength, the smaller the spot) and increasing the numerical aperture of the lens system (that is, the power of the lens) that focuses the laser image; the DVD lens now has a numerical aperture of 0.6.

There are several capacity versions of the DVD disk: a single-layer, single-sided disk at 4.7 GB; a two-sided disk with one layer each side (9.4 GB); a single-sided disk with two layers per side (8.5 GB); and a two-sided disk with two layers each side (17 GB). These disks are all 1.2 mm in thickness but are made with laminates that can be read through from a single side by moving the focus point to the deeper level (two layers), or have to be turned over (two sides). Different types of DVD may also have different capabilities (**Table 2**). The major difference between random access memory (RAM) and read/write (RW) formats is the use of zones in the RAM disk that allow the constant linear velocity operating mode to have a greater degree of rapid random access, closer to a constant angular velocity (CAV) device.

DVD players/recorders are available as stand-alone devices, incorporated into computers, or as part of video entertainment devices or combination devices. Key aspects of DVD players are shown in **Table 3**.

Laser Video disk. Older than DVD technology and now largely supplanted by it, the Laser Video disk

TABLE 2. Types of DVD disks and their uses

Type	Detail	Use
DVD ROM	Read only	Used for DVD playback; not recordable
DVD RAM	Read/write	Recordable and rewritable disk
DVD R	Write once	Recordable once but not re-recordable
DVD RW+/–	Read/write	Different versions of DVD RAM

TABLE 3. Specifications of DVD players

Feature	Specification
Video component	Digital MPEG-2 (30:1 compression from 8-bit 4:2:0)
NTSC resolution	720 × 480 (or PAL resolution 720 × 576)
Audio uncompressed	16/20/24/-bit 48/96 KHz LPCM
Audio compressed	Dolby digital, MPEG1, or MPEG 1 layer 2, audio S/N 96–144 dB
Data rate	Maximum 9.8 Mbps
Rotation speed	570–1600 rpm
Channel data rate	26.16 Mbps

uses many of the same principles. The major difference is that the information is recorded on to the disk and played back in an analog form—that is, it is based on a continuously variable signal without discrete digital bits. The result is less density and less potential for data manipulation, as well as higher cost. There are advantages, however, such as in video quality.

There are three levels of Laser Video disk systems that are commercially available. Level I, the basic level, plays video information in a straightforward linear (CAV or CLV) mode. Movies are typically made at level I. Level I also includes limited interactivity, frame selection, stop and freeze, search options, and chapter stops. Level II includes a microprocessor in the player and has built-in cues in the disk. This allows preprogramming of the video disk with automatic responses and option branches. Level III requires interfacing to a host personal computer. This allows the usual functions of the personal computer to be brought into direct interactivity with the video. Hence the level III system incorporates some adaptive elements of learning into interactivity. *See* COMPUTER; DIGITAL COMPUTER; MICROCOMPUTER.

These functions have resulted in many applications for laser video disks: entertainment; video games with user interaction; instructional devices with branched, cued interactivity; point-of-sale terminals with self-directed pathways; learning systems for self-paced instruction; military training systems; simulation; medical teaching systems; multilanguage training, using dual audio channels; archival storage of pictorial, video, and digital images; and video editing of movies or television. *See* MULTIMEDIA TECHNOLOGY; SIMULATION; VIDEO GAMES.

Video CD. Using the same disk technology used in CD-ROM-XA and combining it with MPEG1 compression, it is possible to play a highly compressed version of a movie. This basically uses 650 MB to record a full movie, and as a result much quality is lost. However, this technology has been very successful in environments where cost is paramount. It is often used for illegal copies of copyright video content. It is not to be confused with an older format called CD-V, which was a hybrid of laser video and CD technology and has largely disappeared. David H. Davies

Bibliography. M. Ely and P. Delancy, *DVD Production with Video*, Butterworth-Heinemann, 2000; A. C. Luther, *Video Recording Technology*, Artech House, 1999; L. Purcell, *CD-R/DVD: Digital Recording to Optical Media*, McGraw-Hill, 2000; J. Ross, *DVD Player Fundamentals*, Prompt Publications, 2000; K. A. Schouhamer Immimk, *Codes for Mass Data Storage Systems*, Shannon Foundation, 1999; J. Taylor, *DVD Demystified*, 2d ed., McGraw-Hill, 2000.

Video games

Entertainment systems in which a computer is used to drive a video display and interact with players using a variety of input devices. Video games can be divided into arcade systems, home computers, and game consoles. The distinction between a home computer and a game console is that a computer can be used for a variety of other applications such as word processing and Internet access while a game console is specifically designed for entertainment purposes.

Arcade systems. Arcade systems are typically built for a particular game or set of related games. Everything from the choice of controls to the design on the sides and face of the unit are geared toward the game itself. The display is typically larger than that on a home computer, providing a greater sense of immersion. In some cases, specialized optics are used to further increase the sense of breadth and depth. In some racing simulations, for example, multiple screens are arranged side by side in an arc around the player in order to give an even greater feeling of being "in" the game.

Input devices on arcade units are more specialized than those on home computers or game consoles. While early arcade systems primarily used

buttons, knobs, and joysticks, the more recent ones have moved far beyond those simple devices. Driving simulations often use full race-car cockpits, and skiing and water-skiing simulators have added a pair of skis that the player can stand on in order to control his or her movements in the game. Light guns which allow the player to "shoot" the screen have proven popular with certain types of action games.

The goal of most of these high-end arcade systems is to provide an experience that cannot be duplicated on a home computer or game console, and without such innovations arcades would rapidly lose business to the growing home market. The chief disadvantage of arcade systems is that they are custom-designed on a per-game basis, which means high development costs and early obsolescence. They are also expensive to distribute, as entire arcade units must be shipped from the manufacturer. There have been at least two initiatives aimed at standardizing arcade hardware so that the units can be reused for different games, but so far there has been relatively little success in that area.

More recently, there has been considerable innovation at the high end of the arcade market. Often described as LBE (location based entertainment), these arcade systems are expensive, custom-designed, and clustered together as part of a larger entertainment complex. These facilities feature a combination of conventional arcade games and specialized high-end systems involving virtual reality gear, large rear-projection screens, motion platforms, and sophisticated input devices. *See* SIMULATION; VIRTUAL REALITY.

Game consoles. A game console typically does not include its own display, but is hooked up to a television set. This provides only limited resolution but dramatically lowers the cost of the consoles themselves. This is critical since a primary selling point of game consoles over home computers is their lower cost. The input devices on game consoles are limited to simple multipurpose controllers that are included with the console. Most consoles come with two controllers in order to support two-player competitive games.

Consoles are much less expensive than the average home computer system and often have specialized hardware for fast graphics and high-quality sound. The games are distributed either as cartridges ("carts") or compact disks (CDs) that contain the game logic programmed into read-only memory (ROM). CD-ROMs have many advantages over carts, including much greater storage capacity and lower production costs. This results in a lower cost of entry to the market, which in turn provides a wider selection of games to the consumer. However, the business model used by many companies that make console games is built on selling the console itself at cost or even taking a small loss on it, and making a significant profit on each game sold. This becomes more difficult if the distribution medium for the game is CD-ROM, since CD-ROM games may be produced by other companies. *See* COMPACT DISK.

Console systems have gone through at least four distinct generations. The first systems were very simple, often involving rudimentary logic that was custom-designed for a particular game. These quickly gave way to systems built around 8-bit microprocessors, such as the original Sega and Nintendo systems. These were superseded by the 16-bit systems which provided greatly increased memory and graphics performance. After some unsuccessful attempts at 32-bit systems, the industry moved to full 64-bit systems, with custom chipsets for sound and three-dimensional (3D) graphics. These systems often rival high-end home computers in terms of graphics capabilities. However, the high cost of market entry for console game developers (relative to the home computer game sector) has limited their selection of games. This, along with the ubiquity of high-performance personal computers in the home, threatens the survival of the console market. *See* BIT; COMPUTER GRAPHICS; COMPUTER STORAGE TECHNOLOGY; MICROPROCESSOR.

Home computers. Home computer systems have grown steadily more powerful to the point where they are more than capable of serving as game machines. Since they use an actual computer monitor rather than relying on a television set, the resolution and overall image quality is much higher than that of game consoles and often rivals or exceeds that of the arcade systems. A variety of controllers are available for the personal computer, ranging from simple analog joysticks to sophisticated input devices with full-force feedback. *See* COMPUTER PERIPHERAL DEVICES.

Production, packaging, and distribution costs for computer games are similar to those for game consoles that use CD-ROM. However, widespread Internet access allows game companies to make sample or "demo" versions of their games available for download. This helps build the market for the games, and may prove to be an effective distribution mechanism for the full retail versions as well. This will avoid the problem of limited shelf-space in computer game stores and will drastically reduce production and distribution costs. *See* INTERNET.

Home computer systems are not without disadvantages. They are typically 10 or 20 times as expensive as an average game console, which limits their market. On the other hand, most families now own a home computer, and many own more than one. Since the computer has its own monitor, the family TV is not monopolized the way it is with a game console.

The other disadvantage that home computers have when compared to arcade systems or game consoles is that there tends to be considerable diversity in terms of hardware capabilities and configuration. Developers must contend with different screen resolutions, color depths, video cards, display drivers, sound cards, input devices, and so on. Even basic information about the end user such as processor speed or the amount of installed memory is unknown during game development. This makes the testing and debugging of computer games much more challenging than game consoles or arcade units.

Types of games. Regardless of the platform (home computer, game console, arcade system), there are a number of basic genres of game. Some of these genres are more common on one type of platform than another, mostly due to technical limitations.

Action games are popular on all platforms. Sometimes called "twitch" games, they tend to be fast-paced and require good reflexes and hand-eye coordination on the part of the player. The storyline for such games is usually very basic—shoot everything that moves. The player is sometimes trying to achieve a goal, such as escaping from a dungeon or rescuing hostages, but the game play is essentially the same regardless.

Adventure games are mostly popular on home computers, though some game consoles do support them. Unlike action games, adventure games combine an element of exploration with some puzzle-solving such as finding keys, unlocking doors, or gathering information. Since these games involve a gradual accumulation of objects in an inventory, as well as the solving of various puzzles, it is essential that there be some way of saving the game state information across playing sessions. On a home computer, this is trivial. Some game consoles also provide some storage capability, although it is usually quite limited. Arcade systems do not have data storage capabilities, so they are generally not suitable for this particular genre. A related genre is role-playing, which is similar to an adventure game but with greater emphasis on playing a particular class of character (such as a magician or a thief) and working to improve that character's skills and abilities.

Simulation games usually involve operating some sort of vehicle, such as an aircraft or a high-performance race car. Flight simulators and vehicle combat simulators (such as tanks or giant walking robots) are in this category as well. Sports simulators are a slightly different category, since they are focused around an individual or team sport such as golf, hockey, football, or soccer. Since this genre shares many characteristics with action games, simulators can also be found on all popular platforms.

Strategy games are based on planning and anticipating future events. There are two subcategories: war games and system simulators. War games are often based on reenacting historic battles of land, sea, or air. The player takes on the role of a general or commander and plans a strategy that hopefully will win the war. System simulators, on the other hand, work by simulating a system such as a city, an ant colony, or an entire planet and allowing the player to make choices that affect the grown and development of that system. Anything from urban planning to maintaining a global ecosystem can be part of the challenge. Since the strategy games often take time to plan and develop, they tend to be found mostly on home computers and are not seen in arcade systems.

Simple puzzle games are the final genre, and they can be found on all platforms. However, their popularity has declined relative to the other gaming genres.

Future directions. A number of new developments will affect arcade systems, game consoles, and home computers. Most important is the steady improvement in the quality and performance of 3D graphics accelerators. Since these chips are designed specifically to handle the final stages of the graphics "pipeline," they can be optimized for that task in order to obtain maximum performance. They also relieve the main central processing unit (CPU) of the burden of doing the low-level graphics rendering, which leaves more computing cycles available for other aspects of the game.

The next area where innovation can be anticipated is the development of 3D audio systems. Just as 3D graphics accelerators take over the work of rendering images on the screen, 3D audio chipsets handle the generation and spatialization of sound. The increasing sophistication of audio hardware and software will allow sound to be positioned in three dimensions around the player as they travel through the game.

The third area that will see rapid growth is support for multiplayer gaming. Home computers already support this feature, using a standard Internet connection to provide head-to-head or team play. Game consoles will likely also gain that capability, and even some arcade units have the ability to link to other systems over custom networks.

Another important development will be the growing use of force feedback, which provides sensations such as vibration, impact, and resistance to movement. Some very high-end arcade systems have even added motion platforms, which tip and tilt the players as they travel through the game.

As the home computer market grows, the other two sectors will continue trying to find ways to compete effectively. New generations of the popular game consoles come out every couple of years, offering graphics and sound improvements that overshadow all but the highest-end home computer systems. Arcade units are attempting to differentiate themselves from both home computers and game consoles by providing ever more sophisticated input devices. This competition between platforms is likely to produce some amazingly sophisticated and compelling games. *See* HUMAN-COMPUTER INTERACTION.

Bernie Roehl

Bibliography. *Computer Gaming World Magazine*; *Game Developer Magazine*; *Next Generation Magazine*.

Video microscopy

The use of a high-quality video camera or other fast camera (such as a charge-coupled device) attached to a research-quality light microscope for the purpose of real-time or high-speed imaging of samples on a microscope stage. These images are recorded at regular intervals (often at "video rate" of 30 images per second), and the time-lapse sequence can be played back in the form of a movie. The term "video microscopy" originally referred to microscope imaging

using true video (30 frames per second) but now generally refers to rapid time-lapse imaging techniques. Video microscopy is used frequently to image small structures that move rapidly within cells as well as movement of whole cells. This motion can be quantitated, and in the case of fluorescence microscopy, changes in fluorescent intensity (reflecting the local chemical environment of the fluorescent molecule or the number of fluorescent molecules) can be quantitated as well. *See* CAMERA; MICROSCOPE.

Conventional video systems. Video microscopy is not very useful for stained biological material or specimens that have high contrast. It is most useful for specimens that have extremely weak contrast or are invisible by conventional microscopy. In practice, the video signal from the camera is digitized and enhanced at video frequency by an image processor. Contrast is enhanced first electronically by changing the offset of the camera, and then digitally by choosing the peak gray levels (usually about 30 levels) of the biological image and expanding them over 256 levels of gray. The latter operation is performed by assigning each gray level in the original image to a new value through the use of an output look-up table. *See* OPTICAL MICROSCOPE.

Image processors are able to perform operations in real time. The image processor is also used to improve image quality by averaging successive frames to remove electronic noise and by subtracting a stored image of the background to remove fixed-pattern mottle (such as dust spots) on the lenses. Subtraction of the background image at video frequency also serves to correct minor unevenness of illumination. The resulting image exhibits high contrast and additional detail that is invisible when viewed by the eye through the microscope. Microscopes equipped with Nomarski differential interference contrast optics offer the best visualization of submicroscopic objects by video microscopy. This type of microscopy permits visualization of biological objects as small as 20 nanometers and solid objects as small as 5 nm (for example, colloidal gold particles). Biological objects detected by this method are an order of magnitude smaller than the limit of resolution of light microscopes. *See* IMAGE PROCESSING; INTERFERENCE MICROSCOPE.

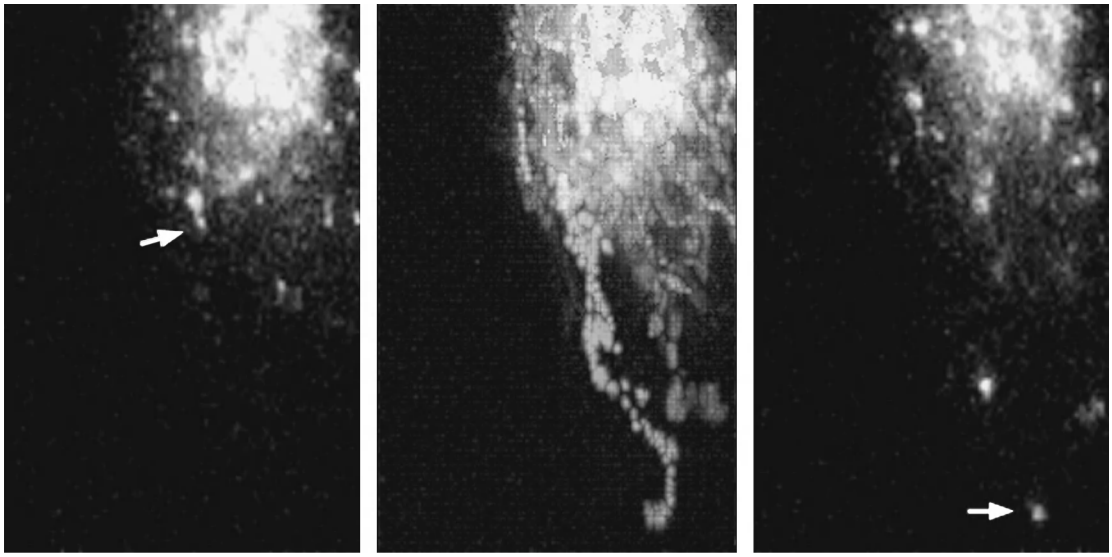
Charge-coupled-device systems. Charge-coupled-device (CCD) cameras use a light-sensitive silicon chip as the detector. Often the chip is cooled to very low temperatures (-20 to -40°C) to reduce noise. CCD chips have the advantage that the signal can be easily digitized and stored on a computer, and this digital signal is a linear measure of the amount of light recorded by the chip (and therefore of the number of fluorescent molecules). They are often better than conventional video cameras at recording images of fluorescent objects (since fluorescence images are often dim, and CCD cameras allow varying the length of exposure to collect more signal). Conventional video cameras are still preferred if transmitted light is used. *See* CHARGE-COUPLED DEVICES.

Motility assays. Video microscopy is used routinely to examine the structure and function of the cellular cytoskeleton and to detect the movement of submicroscopic organelles in living cells. The filaments of the cytoskeleton function as tracks on which protein "molecular motors" transport material to different locations in the cell. Microtubules (one class of cytoskeletal filaments) and membrane-enclosed vesicles, both smaller than the limit of resolution, can be detected by video microscopy in living cells and in reconstituted cell-free systems. *See* CYTOSKELETON.

The video microscope has also been used to examine the movement of objects by molecular motors during motility assays. The most common type of motility assay is the gliding filament assay. First, the motor protein of interest is adsorbed to a glass coverslip. Then the cytoskeletal filaments on which these motors move are added to the coverslip, and video microscopy is used to view the movement of filaments over the glass surface. The filaments appear to glide over the glass surface. This assay has become a standard procedure for studying the microtubule-dependent motors, kinesin and dynein. *See* CELL MOTILITY.

Another method is to adsorb the filaments to the glass surface first, then add organelles or artificial beads coated with motors, and view movement of the organelles and beads along the filaments. The latter method has been used successfully to study both microtubule-dependent motors and the actin filament-dependent motors known as myosins. Actin filaments (another class of cytoskeletal filaments) are too small to be detected by video microscopy, but vesicles moving on them can be seen.

Fluorescence ratio imaging. Rapid advances in low-light-level cameras have made it possible to expand video microscopy to study activities in living cells by using fluorescent dyes. Living cells can be stained with an ion-sensitive fluorescent dye, and changes in the fluorescence excitation or emission can be used to monitor cellular activity. The quantitation of the optical signal from living cells is best achieved by using fluorescence ratio imaging. This technique is used to measure ion concentrations such as hydrogen and calcium, and requires that the fluorescent probe be sensitive to changes in the concentration of the ion of interest at one excitation or emission wavelength but insensitive at another. The ratio of the fluorescence intensity at the two different wavelengths normalizes for optical pathlength, local probe concentration, and loss of signal due to photobleaching (loss of fluorescence due to exposure to light). The fluorescence ratio-imaging microscope requires a standard fluorescence microscope, a low-light-level detector (often a CCD camera), and an image processor. An image processor is required for image acquisition and basic arithmetic operations such as addition and subtraction; basic point operations such as intensity remapping and thresholding; basic analysis operations such as histogram and statistical calculations; and basic display operations such as image scroll, zoom, and pseudocolor display. *See* FLUORESCENCE; FLUORESCENCE MICROSCOPE.



Use of video microscopy to track an intracellular object (at arrow) labeled with a GFP-tagged protein as it moves away from the Golgi apparatus (large area of bright fluorescence). The beginning frame (left) and ending frame (right) of the sequence are shown. The center frame is a superposition of the entire sequence showing the object's path as it moves through the cell.

Fluorescence recovery after photobleaching. Video microscopy in combination with fluorescent probes can also be used to measure dynamic changes in the structural components in cells or the rate of diffusion of macromolecules in cells. Since most fluorescent probes bleach when illuminated for some period at the excitation wavelength, this property can be exploited to measure dynamic changes in cells. This technique is referred to as fluorescence recovery after photobleaching, and requires that the area of the cell around the region of interest be masked to protect it from the light while the region of interest is illuminated. Fluorescent molecules from the unbleached region diffuse to the bleached region, and the rate of recovery of fluorescence can be determined. This technique has been used to demonstrate that cytoskeletal filaments exchange subunits with a monomer pool and that macromolecules diffuse slowly through the porous matrix of the cytoplasm. The technique has also been used to determine how fast the membrane proteins diffuse in membranes such as the plasma membrane. Importantly, it is easy to distinguish which molecules are fixed in place in the cell (by attachment to cytoskeleton or other structures) and which are free to diffuse. *See* CELL ORGANIZATION.

Green fluorescent protein. Until recently, attaching fluorescent probes to proteins required either chemical attachment of the fluorescent probe to the protein of interest followed by introduction of the protein to the cell, or the addition of fluorescently labeled antibodies to the protein of interest in situ. These methods work well if the protein of interest is one that is normally taken up by the cell or if it is exposed on the cell surface. However, the overwhelming majority of cellular proteins could be fluorescently labeled only with great difficulty if they were to be used in a living cell.

The green fluorescent protein (GFP) from the jellyfish *Aequorea victoria* has been used as a fluorescent tag. GFP is fluorescent without the attachment of another molecule. Well-established recombinant DNA techniques make it possible to make fusion proteins in which GFP is attached to any cellular protein. Upon introduction of fusion protein DNA into cells in culture, the protein tagged with GFP is synthesized by the cell just like the normal cellular version of the protein. Since this protein is fluorescent, its movements inside the cell can be easily followed by video microscopy. The **illustration** shows an example of a structure visualized with a GFP-tagged protein moving within a cell. It is important to verify that the GFP tag does not interfere with the normal function of the protein. GFP variants (and even a red fluorescent protein) have been developed that allow for tagging and following up of three proteins simultaneously in the same cell. *See* BIOLUMINESCENCE; GENETIC ENGINEERING.

George M. Langford; John F. Presley

Bibliography. R. D. Allen, The microtubule as an intracellular engine, *Sci. Amer.*, 255(2):42–49, 1987; R. D. Allen and N. S. Ailen, Video-enhanced microscopy with a computer frame memory, *J. Microsc.*, 129:3–17, 1983; S. Inoue and K. Spring, *Video Microscopy: The Fundamentals*, 1997; S. Scales, R. Pepperkok, and T. E. Kreis, Visualization of ER-to-Golgi transport in living cells reveals a sequential mode of action for COPII and COPI, *Cell*, 90:1137–1148, 1997; D. L. Taylor and Y.-L. Wang (eds.), *Methods in Cell Biology*, vol. 30, 1989.

Videotelephony

A means of simultaneous, two-way communication comprising both audio and video elements. Participants in a video telephone call can both see and hear each other in real time. Videotelephony is a subset of teleconferencing, broadly defined as the various

ways and means by which people communicate with one another over some distance. Initially conceived as an extension to the telephone, videotelephony is now possible using computers with network connections. In addition to general personal use, there are specific professional applications, such as criminal justice, health care delivery, and surveillance that can greatly benefit from videotelephony. *See* TELECONFERENCING.

History. Although basic research on the technology of videotelephony dates back to 1925, the first public demonstration of the concept was by American Telephone and Telegraph Corporation (AT&T) at the 1964 New York World's Fair. The device was called Picturephone, and the high cost of the analog circuits required to support it made it very expensive and thus unsuitable for the market place. The 1970s brought the first attempts at digitization of transmissions. The video telephones comprised four parts: a standard touch-tone telephone, a small screen with a camera and loudspeaker, a control pad with user controls and a microphone, and a service unit (coder-decoder, or CODEC) that converted the analog signals to digital signals for transmission and vice versa for reception and display. This system required a wideband infrastructure (telephone circuits) with a capacity of 6.3 megabits/s. *See* ANALOG-TO-DIGITAL CONVERTER; DIGITAL-TO-ANALOG CONVERTER.

During the 1980s companies attempted to market videoconferencing for both public and private use, for example, setting up nationwide networks of videoconference rooms.

Analog video transmission. The primary limiting factor in videotelephony development has been the requirement for switched broadband circuits. The public telephone infrastructure was developed for narrow-band voice communication only. Television network broadcasts and television cable distribution systems are wideband in nature, but they are not switched between users and are only one-way. Transmission requirements can be reduced considerably by utilizing smaller screens with less spatial resolution, by reducing image refresh rates, and by delivering black-and-white images only. However, the requirements are still far beyond the capacity of ordinary telephone lines. Due to these limitations, true analog videotelephony has never really developed. Therefore, considerable attention has been devoted to digitizing and compressing video signals to make their transmission more economical. *See* CABLE TELEVISION SYSTEM; TELEPHONE SERVICE; TELEVISION NETWORKS.

Digital video transmission. Pure digitization of video signals is normally done using pulse code modulation (PCM), which requires sampling and encoding to produce a transmission rate of 45 megabits/s. The video telephones from the 1970s used the differential pulse-code modulation (DPCM) compression technique, which reduced the required transmission rate to 6.3 megabits/s. This technique transmits only the differences between successive samples. Aother compression technique popularized in the 1980s is the discrete cosine transform (DCT), which takes advantage of the fact that several adjacent points within an image essentially are the same. A third, widely used compression algorithm, called motion compensation, is based on the fact that moving objects keep the same form. Thus an image of an object must be transmitted only once, and its subsequent movement requires transmission of only its new position and orientation. *See* DATA COMPRESSION; PULSE MODULATION.

During the 1980s the International Telecommunications Union (ITU) agreed to use a standard for video coding and decoding called H.261. This standard uses a combination of the three compression algorithms and specifies standard image formats. Using the standard, good-quality video images may be communicated at transmission rates as low as 64 kilobits/s. A slightly different version of this standard is also common, called H.263. Several audio standards are in use: G.723, G.711, G.722, and G.728. Two general, higher-level videotelephony/videoconferencing standards developed by the ITU are commonly adhered to: H.320 for ISDN videotelephony and H.323 for IP videotelephony.

Infrastructure. Considerable improvements in telephone, computer, and television infrastructures have been made over the years. Telephone connections over copper cable have been replaced by optical fiber, the capacity of connections to the home has been increased and standardized, and more efficient technologies to use transmission capacity are available. Cable television connections have become interactive. Computer/data communication infrastructures, such as the Internet, World Wide Web, and high-speed, wide-band links such as digital subscriber line (DSL) and cable modems have become more widespread. These developments have converged to make video communication possible at affordable rates. *See* INTERNET; MODEM; OPTICAL COMMUNICATIONS; WORLD WIDE WEB.

The basis of communication over the Internet is the Internet Protocol (IP), which allows for data in packets to be routed to different addresses using the Transmission Control Protocol (TCP). The data in the packets may be voice, video, data, graphics, and so forth. Thus, multimedia exchanges may take place between two or more parties. The exchanges can be over the public Internet or over private corporate intranets. Some companies have started replacing their traditional telephone private branch exchanges (PBXs) with IP-based PBXs, which would theoretically allow everyone in the company to have a video or multimedia telephone—totally separate from a computer. This could provide, for example, unified messaging—the ability to store voice mail as e-mail, answer telephone calls with e-mails and vice versa. *See* ELECTRONIC MAIL; PRIVATE BRANCH EXCHANGE.

Video telephone hardware. Small residential video telephones, computer-based desktop video telephones, and small videoconferencing setups have been introduced to fulfill diverse needs. One such

commercially available residential videophone is about as big as a typical office desk telephone with a small flip-up screen that has an eyeball camera above it. Although it will work with several standards, this phone is primarily designed for use over Integrated Services Digital Network (ISDN) lines in which a residence gets three circuits; one circuit is used for control and the other two for voice and video. *See* INTEGRATED SERVICES DIGITAL NETWORK (ISDN).

An example of a computer-based desktop videophone consists of a PCI (Peripheral Component Interconnect) video/audio CODEC board to add to a personal computer, a composite color camera, audio peripherals, and visual collaboration software.

Video telephone software. Videotelephony software has been developed and made widely available that permits real-time collaboration and conferencing, including multipoint and point-to-point conferencing. Multipoint means, for example, that three people in three different locations could have a video telephone conference call in which each could see and hear the others. In addition to the basic audio and video capabilities, such software provides several other features such as a whiteboard, background file transfer, program sharing, and remote desktop sharing.

John Bleiweis

Bibliography. E. M. Dickson and R. Bowers, *The Video Telephone*, 1974; S. J. Emmott and D. Travis (eds.), *Information Superhighways: Multimedia Users and Futures*, 1995; K. Finn, A. J. Sellen, and S. B. Wilbur, *Video-Mediated Communication* (*Computers, Cognition and Work*), April 1997; A. M. Noll, Anatomy of a failure: Picturephone revisited, *Telecomm. Policy*, 16:307–316, 1992; R. Schaphorst, *Videoconferencing and Videotelephony: technology and Standards*, Artech House Telecommunications Library, January 1997.

Vinegar

A food condiment containing mainly acetic acid that is produced by the bacterial oxidative fermentation of various ethanolic solutions. Vinegar, like wine and certain milk products, is one of the oldest fermented foods used by humans. Babylonian records indicate that vinegars prepared from wines and beer were widely used as early as 5000 B.C. *See* ACETIC ACID; WINE.

Types. The names for the different vinegars are based on the substrates from which they are made. These include the juices from different fruits, starchy vegetables, cereals, and distilled ethanol (ethyl alcohol). In the United States, white distilled vinegar makes up over 80% of the annual production. *See* ETHYL ALCOHOL.

Much of the white distilled vinegar is made from denatured synthetic ethanol. Nitrogen compounds, minerals, and other nutrients must be added to the alcohol medium to support the growth and metabolism of the acetic acid bacteria. Acetic acid or ethyl acetate are commonly used as the ethanol denaturants.

Most other vinegars are made from ethanolic solutions generated by a yeast fermentation. Fruit juices usually contain sufficient concentrations of hexose sugars and other yeast nutrients to support yeast growth and fermentation. Cider vinegar is produced from yeast-fermented apple juice, wine vinegar from fermented grape juice. Other juices such as pineapple, orange, and pear, as well as sugarcane syrup and molasses, can serve as fermentation substrates.

Vinegar is also produced from starchy vegetables such as potatoes and from cereals such as barley, corn, and rice. Yeasts do not ferment starch to ethanol, and therefore the starch first must be broken down (saccharified) to simple sugars, mainly glucose. Malt vinegar is made from an infusion of barley malt and other cereals in which the enzymes in the malt have converted the starch to fermentable sugars. *See* ENZYME.

Balsamic vinegar from Modena, Italy, is a costly product that is highly prized by gourmets. The juice of Trebbiano grapes is concentrated about 30% by simmering, then fermented by *Zygosaccharomyces* yeasts that tolerate high concentrations of sugar. A portion of the ferment is transferred annually to a succession of wooden casks, where the acetic acid fermentation and other enzymatic changes occur. The battery of casks may be 200 years or older; thus, the harvested vinegar is of great age. About 2 liters (2 qt) of vinegar are produced from 100 kg (220 lb) of the concentrated grape must.

Vinegar sold in the United States must contain at least 4 g of acetic acid per 100 ml of solution

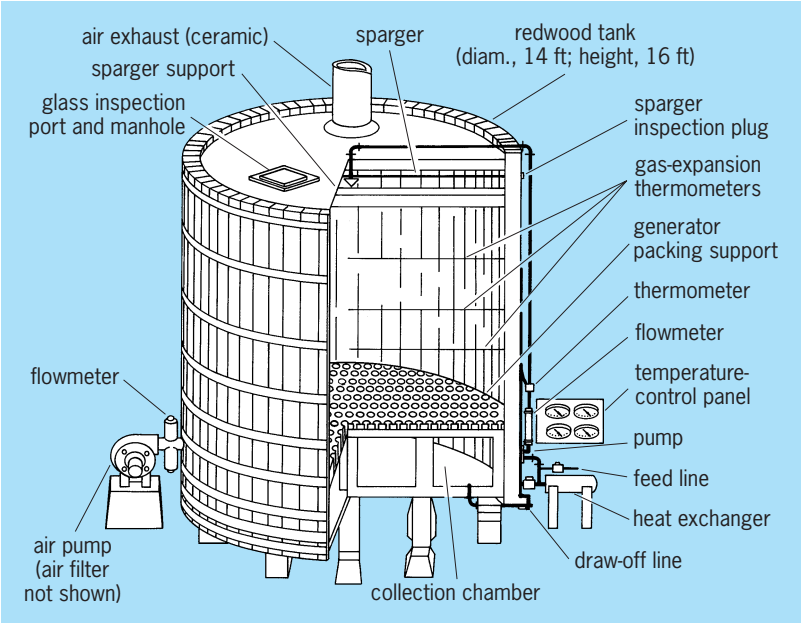


Fig. 1. Trickling, quick vinegar generator. The diameter of the redwood tank is 14 ft (4 m), and the height is 16 ft (5 m). (*After L. A. Underkofler and R. J. Hickey, eds., Industrial Fermentations, 2 vols., Chemical Publishing, 1954*)

(40-grain vinegar). Other federal specifications describe permitted color, odor, presence of trace metals, and alcohol content. Nondistilled vinegars possess distinctive colors and flavors that reflect the properties of the original substrate.

Acetic acid bacteria. Bacteria belonging to the genus *Acetobacter* are primarily responsible for the vinegar fermentation. Four species are recognized; they are distinguished on the basis of nutrient requirements, production of brown pigments, and tolerance to ethanol. All acetic acid bacteria require oxygen for growth and metabolism.

Gluconobacter is a closely related genus that also oxidizes ethanol to acetic acid. The genus is differentiated from *Acetobacter* by the fact that it cannot oxidize acetic acid to carbon dioxide and water. *Acetobacter* species carry out the latter oxidation, often after ethanol has been depleted. *See* INDUSTRIAL MICROBIOLOGY.

Production methods. A variety of methods are used for the production of vinegar. The simplest fermentations are slow and inefficient but are relatively foolproof and utilize inexpensive equipment. Submerged fermenters produce 120–150-grain vinegar in minimal time, but they are expensive and system failures can occur.

Orleans process. This is one of the older commercial methods. The fermentation is carried out in a large cask in which holes have been drilled to permit the introduction of air. Screens are used to prevent the entrance of fruit flies. The cask has a spigot for the withdrawal of finished vinegar, and the bung (stopper) contains a tube so that fresh wine or other substrate can be added without disturbing the film of vinegar bacteria. The fermentation is initiated by mixing wine with about 20% fresh vinegar. Although the process is slow and the yields of acetic acid may be under 80%, vinegar of high quality can be produced.

Trickling, quick process. In this procedure, the alcoholic substrate is sprayed on the surface of a chamber packed with beechwood shavings, corn cobs, or other carrier materials that support a slime composed of acetic acid bacteria (**Fig. 1**). Air is sparged through the system. Because considerable heat is generated, cooling of the chamber or recirculated vinegar stock is usually required. The vinegar stock is recirculated two or three times until the desired concentration of acetic acid is attained. Generators of this type may be operated continuously for several years or more.

Submerged fermentation. In submerged fermentation (**Fig. 2**), air is sucked into the bottom of a stainless steel tank and then broken into very small bubbles by the action of a high-speed rotor. Advantages of this system are that vinegar is produced at a high rate, about 10-fold faster than the trickling filter, and the efficiency may be as high as 94%. One disadvantage is that loss of aeration for even 1 min may result in death of the culture. Because of this problem, many vinegar producers have auxiliary electric generators as backups in the event of power failure.

Final processing steps are filtration to clarify the vinegar and pasteurization at 140–158°F (60–70°C) to arrest further bacterial growth and enzyme activity. When concentrated vinegar is desired, for example, 200-grain, this can be accomplished by freezing out some of the water from a 120-grain product. *See* FERMENTATION.

Don F. Splittstoesser

Bibliography. H. A. Conner and R. J. Allgeier, Vinegar: Its history and development, *Adv. Appl. Microbiol.*, 20:81–133, 1976; H. J. Peppler and D. Perlman (eds.), *Microbial Technology*, 2d ed., vol. 2, 1979; M. Plessi, A. Monzani, and D. Coppini, Determination of the monosaccharide and alcohol content of balsamic and other vinegars by enzymatic methods, *Agr. Biol. Chem.*, 52:25–30, 1988; H. J. Rehm and G. Reed (eds.), *Biotechnology*, vol. 3, 2d ed., 1994.

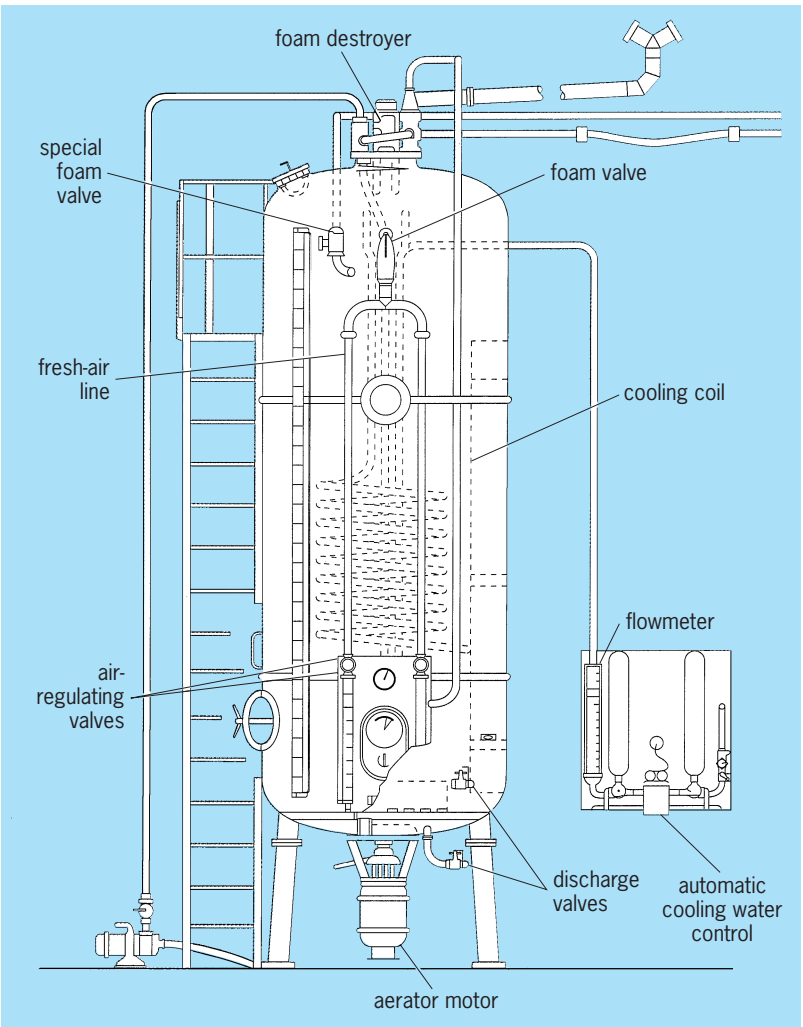


Fig. 2. Frings acetator for submerged fermentation. (*After G. B. Nickol, Microbial Technology, 2d ed., vol. 2, Academic Press, 1979*)

Violales

An order of flowering plants, division Magnoliophyta (Angiospermae), in the subclass Dilleniidae of the class Magnoliopsida (dicotyledons). The order

consists of 24 families and more than 5000 species, including most of the families that have been referred to the Englerian order Parietales. The largest families are the Flacourtiaceae (about 800 species), Violaceae (about 800 species), Cucurbitaceae (about 900 species), Begoniaceae (about 1000 species), and Passifloraceae (about 650 species). The Cucurbitaceae have often been considered to constitute a distinct order, Cucurbitales, but there is a developing consensus that their relationship is here.

The most characteristic feature of this morphologically heterogeneous order is the unilocular, compound ovary with mostly parietal placentation. When the stamens are numerous they are initiated in centrifugal sequence. The ovules have two integuments and a nucellus that is several cells thick. The more primitive members of the order, such as some of the Flacourtiaceae, are trees with stipulate, alternate leaves; perfect, polypetalous flowers with numerous centrifugal stamens; a compound pistil with free styles and parietal placentation; and seeds with a well-developed endosperm. Tendencies toward unisexuality, reduction in the number of stamens, fusion of filaments, development of a corona, reduction in the number of carpels, fusion of styles, and loss of endosperm from the seed can all be seen in the Flacourtiaceae. These are also some of the prominent characteristics used in combination to define many of the other families of the order. *See* FLOWER; LEAF; SEED.

The Violales differ from the related order Capparales (which also has parietal placentation) in usually lacking myrosin cells, in having a much higher proportion of woody species, in often having perigynous to epigynous flowers, in seldom having compound leaves, and in very often having three carpels (a rare number in the Capparales) and only seldom two (the most common number in the Capparales). *See* CAPPARALES.

Cucumbers and melons (species of *Cucumis*, in the Cucurbitaceae), pumpkins and squashes (species of *Cucurbita*), begonias, and violets (*Viola*; see **illus.**) are familiar members of the Violales. *See* CANTALOUPE; CUCUMBER; DILLENIIDAE; HONEYDEW MELON; MAGNOLIOPHYTA; MAGNOLIOPSIDA; MUSKMELON; PERSIAN MELON; PLANT TAXONOMY; PUMPKIN; SQUASH; WATERMELON.

Arthur Cronquist; T. M. Barkley

Downy yellow violet (*Viola pubescens*). (*Photograph by Elsie M. Rodgers, National Audubon Society*)

Viral inclusion bodies

Abnormal structures formed in the cytoplasm or the nucleus of the host cell (or both) during the multiplication of some viruses. These structures vary from virus to virus and may represent aggregations of mature virus particles but more often are areas where viral synthesis takes place. These areas of viral synthesis have an altered appearance when treated with typical diagnostic stains.

The presence of inclusion bodies is often important in diagnosis. Brain tissue taken from suspected cases of rabies is examined for the presence of cytoplasmic inclusions known as Negri bodies. In papillomavirus infections, inclusion bodies are found in the nuclei of the epidermal cells in which typical virus particles can be seen by electron microscopy. In human herpes simplex infection, intranuclear inclusions are found that are surrounded by a clear halo. These inclusions are sites of replication of the virus, or "virus factories." *See* HERPES; RABIES; VIRUS.

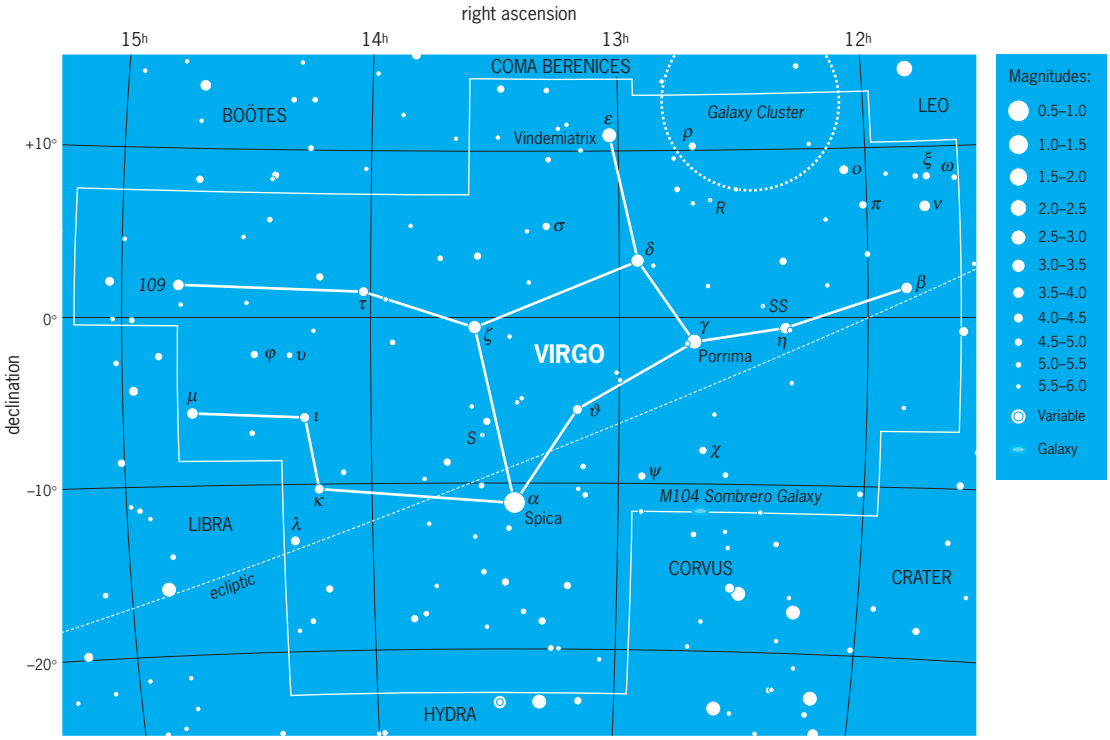
Marcia M. Pierce

Bibliography. L. Collier and J. Oxford, *Human Virology*, 2d ed., Oxford University Press, Oxford, U.K., 2000; D. M. Knight and P. M. Howley (eds.), *Fundamental Virology*, 4th ed., Lippincott, Williams & Wilkins, Philadelphia, 2001; L. M. Prescott, J. P. Harley, and D. A. Klein, *Microbiology*, 6th ed., McGraw-Hill, New York, 2005; S. Schlesinger and M. Schlesinger, Viruses, in J. Lederberg (ed.), *Encyclopedia of microbiology*, 2d ed., vol. 4, pp. 796–810, Academic Press, San Diego, 2000; B. A. A. Voyles, *The Biology of Viruses*, 2d ed., McGraw-Hill, Chicago, 2002.

Virgo

The Virgin, perhaps Astraea, in Greek mythology the goddess of justice, whose scales (the constellation Libra) are adjacent. Also, she could be Ceres, goddess of agriculture, or her daughter Proserpina who, because she ate from a forbidden pomegranate, must spend half her year with Pluto, her husband, in Hades, leaving winter behind for the rest of us. This southern zodiacal constellation is visible in the summer evening sky from midnorthern latitudes. It contains the nearest cluster of galaxies to us; the Milky Way Galaxy is in an outer part of the cluster. *See* LIBRA; VIRGO CLUSTER; ZODIAC.

Spica, the brightest star in Virgo, at first magnitude is the 16th brightest star in the sky. It is a B-star, a hot star on the main sequence of the Hertzsprung-Russell diagram, and glows blue-white. It is slightly variable in brightness. Since Spica means an ear of wheat or corn, it would match the identification of Virgo with Ceres or Proserpina. *See* SPICA.



Modern boundaries of the constellation Virgo, the Virgin. The celestial equator is 0° of declination, which corresponds to celestial latitude. Right ascension corresponds to celestial longitude, with each hour of right ascension representing 15° of arc. Apparent brightness of stars is shown with dot sizes to illustrate the magnitude scale, where the brightest stars in the sky are 0th magnitude or brighter and the faintest stars that can be seen with the unaided eye at a dark site are 6th magnitude. (*Wil Tirion*)

The modern boundaries of the 88 constellations, including this one, were defined by the International Astronomical Union in 1928. *See* CONSTELLATION.

Jay M. Pasachoff

Virgo Cluster

The nearest large cluster of galaxies and one of the best studied. It dominates the distribution of bright galaxies in the sky, forming a conspicuous clump in the distribution of easily visible elliptical and spiral galaxies. The clump lies mostly in the constellation of Virgo but also extends into neighboring constellations, especially Coma Berenices. The well-known spirals M58, M61, and M100, for example, are members of the Virgo Cluster. *See* CONSTELLATION; UNIVERSE; VIRGO.

Structure. The first exhaustive exploration of the Virgo Cluster was the Du Pont Telescope survey, carried out in the 1980s by A. Sandage and B. Binggeli. Their survey showed that the cluster has the appearance of two superimposed clusters (**Fig. 1**). The two concentrations are approximately 5° apart; one is centered near the very luminous elliptical galaxy M87 and the other near M49, another giant elliptical galaxy. They lie along a nearly north–south line. The spiral galaxies are much less concentrated toward these central peaks than are either the giant ellipticals or the dwarf ellipticals. Furthermore, the spirals are more common, relative to other types, in the southern concentration than in the northern. It is also found that the dwarf elliptical galaxies are less concentrated toward the center of the entire complex than are any of the other types, possibly the result of mass segregation, with the more massive

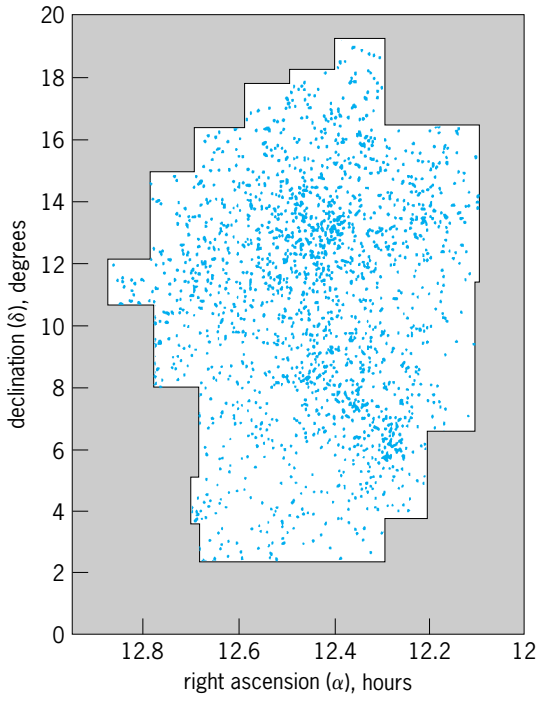


Fig. 1. Distribution of members of the Virgo Cluster on the sky, as determined in the Du Pont Telescope survey. (*After O.-G. Richter and B. Binggeli, eds., The Virgo Cluster, European Southern Observatory, 1985*)

galaxies, under the effects of gravity and encounters, having fallen toward the center of mass more than the low-mass ones.

Populations. The Virgo Cluster is an example of a mixed-population cluster, with both early-type (elliptical and S0) and late-type (spiral and irregular) galaxies present. For members brighter than apparent blue magnitudes of approximately 15.7, about 75% are early types; most of these are dwarf galaxies of low intrinsic luminosity. The remainder are mostly spiral and dwarf irregulars. The most common type of spiral is Sc, of which there are 32 members out of the whole sample of 1343 probable members brighter than this limit (**Fig. 2**).

The total population of all types and of all intrinsic luminosities is not known, but it can be estimated from the surveys and by extrapolating to the faint end of the galaxy luminosity function (the relationship between number and luminosity of galaxies). For galaxies brighter than an absolute magnitude of about −10, a total population of 1900 galaxies is calculated. This excludes the several outlying groups and clusters, such as the Local Group (which includes 30 galaxies). *See* LOCAL GROUP.

Galaxy interactions. As the nearest large cluster, the Virgo Cluster provides a good testing ground for theories of how the dense cluster environment affects the properties of galaxies. The first peculiarity of the Virgo galaxies to be noticed was the rather limited extent of the disks of the spirals. Especially evident in their distribution of neutral hydrogen, as measured at radio wavelengths, the spiral galaxies clearly have been stripped of their outer gas by encounters with other galaxies in the cluster. A related phenomenon is the presence of so-called anemic galaxies, spirals that have unusually low surface brightness and inconspicuous spiral arms. There are several such galaxies in the Virgo Cluster, the result of direct collisions or other tidal disturbances that have removed gas from the disks, inhibiting star formation there and leaving unusually pale galaxies in place of once-robust ones.

Recent studies with large optical telescopes (for instance, with the Keck Telescopes, the Very Large Telescope, the Gemini North Telescope, and other giant telescopes), together with infrared, x-ray, and radio investigations, have shown evidence of violent gravitational interactions among members. For instance, the galaxy NGC 4438, noted as peculiar years ago from Palomar Observatory photos, has been found to have a long tidal tail of debris (stars and gas), ejected by a near-collision with a neighbor. A large cloud of gas lies to one side, produced by ram-pressure stripping, apparently in an earlier collision.

A massive cloud of very hot gas that surrounds the Virgo Cluster is detected by x-ray and radio telescopes (**Fig. 3**). Apparently it is the result of the heating of both intergalactic and intragalactic gas, some of it ejected by collisions of galaxies.

Surveys from space. Surveys of the Virgo Cluster galaxies have used various large ground-based and space telescopes to examine the members in detail. Notable results of some of these surveys have shown

Fig. 2. Central area of the Virgo Cluster, showing some of its brightest galaxy members. (*Canada-France-Hawaii Telescope image, courtesy of C. Cuillandre*)

dramatic evidence of the complicated evolution that occurs in a cluster environment. There is a large number of dwarf elliptical galaxies, many with bright nuclei, which are found to be remnants of low-mass star-forming galaxies that were stripped of their gas in the recent past. Massive galaxies like M87 (**Fig. 4**) are found to have captured intergalactic gas, which has fallen toward its nucleus, where it is swallowed by a massive black hole. Just before it disappears, the gas is violently heated and some is ejected as a jet that can be detected at virtually all wavelengths, from x-rays to radio waves. Surveys by the *Hubble Space Telescope*, the *Spitzer Infrared Telescope* (**Fig. 5**), and the *Chandra X-ray Observatory* have brought new understanding of galaxy evolution in the cluster, where interactions and collisions leave their

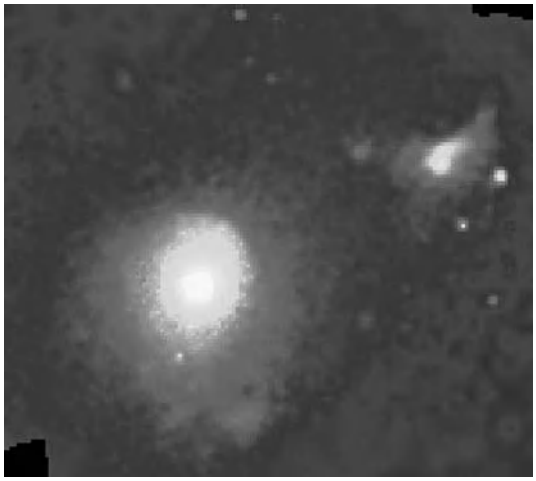


Fig. 3. X-ray image of the entire Virgo Cluster. The main blob is centered on the giant active galaxy M87 (see Fig. 4). (*ROSAT image, courtesy of NASA*)

Fig. 4. Giant elliptical galaxy M87. It is one of the brightest galaxies in the Virgo Cluster and is notable for its extremely massive central black hole. (*CFHT image, courtesy of C. Coullandre*)

indelible marks. *See* CHANDRA X-RAY OBSERVATORY; HUBBLE SPACE TELESCOPE; SPITZER SPACE TELESCOPE.

Dynamics. The dynamics of the cluster is more complex than was once thought. It is not a simple, single group, acting under gravity in a symmetrical way. The velocities in the core are complicated. The elliptical galaxies form a relatively regular, concentrated subcluster that may constitute a dynamically smoothed system, in which gravitational interactions have erased any irregularity of shape. However, the spirals have a wider distribution in space and an accompanying larger spread in velocities; they seem to be still in the process of falling into the cluster core.

The velocities can give a value for the entire cluster mass and for the ratio of mass to light, a useful ratio for gauging the amount of dark matter in a cluster. For the Virgo Cluster, the mass-to-light ratio is found to be in the range of 500 in solar units, depending on how well smoothed the velocities have become, which is still uncertain. This value indicates that the cluster contains mostly unseen matter, as the average mass-to-light ratio for a population of normal stars is about 1. The total mass of the cluster is calculated to be 4×10^{14} times the mass of the Sun. *See* COSMOLOGY.

Distance. Determining the distance to the Virgo Cluster has been one of the more important tasks of twentieth-century astronomy. The Virgo Cluster forms an important stepping-stone for the cosmic distance scale, as it provides a connection between local distance criteria and the use of the Hubble law of gauging distances. The best measurements of the distance to the Virgo Cluster are based on the study of its cepheid variables with the *Hubble Space Telescope*. Other evidence comes from study of its globular star clusters, its planetary nebulae, its novae, and its supernovae. These give a mean distance of 16 megaparsecs (5×10^7 light-years). When the infall velocity of the Milky Way Galaxy toward the center of the Virgo Cluster is taken into account, this distance indicates that the local cosmic value of the Hubble constant, which relates expansion velocity to distance, is about 70 kilometers per second

Fig. 5. Sombrero Galaxy, in an image obtained with the *Spitzer Infrared Orbiting Telescope*. The spheroidal galaxy of stars hosts a thin, spiral-shaped ring of dust, visible in the infrared because of the light emission from the warm dust grains. (*NASA Spitzer image*)

per megaparsec. *See* CEPHEIDS; GALAXY, EXTERNAL; HUBBLE CONSTANT. Paul Hodge

Bibliography. A. Diaferio and K. van der Huch (eds.), *Outskirts of Galaxy Clusters*, Cambridge University Press, 2004; O.-G. Richter and B. Binggeli (eds.), *The Virgo Cluster*, European Southern Observatory, 1985; W. C. Saslaw, *The Distribution of Galaxies: Galaxy Clustering in Cosmology*, Cambridge University Press, 2000; W. Waller and P. Hodge, *Galaxies and the Cosmic Frontier*, Harvard University Press, 2003.

Virial equation

An equation of state of gases that has additional terms beyond that for an ideal gas which account for the interactions between the molecules. The pressure p can be expressed in terms of the molar volume $V_m = V/n$ (where n is the number of moles of gas molecules in a volume V), the absolute temperature T, and the universal gas constant $R = 8.3145$ J K^{-1} mol^{-1} or, in more commonly used practical units, 0.082058 L atm K^{-1} mol^{-1} [Eq. (1)]. In the equation

$$\frac{V_m}{RT} = 1 + \frac{B_2(T)}{V_m} + \frac{B_3(T)}{V_m^2} + \frac{B_4(T)}{V_m^3} + \cdots \qquad (1)$$

the virial coefficients $B_n(T)$ are functions only of the temperature and depend on the nature of the gas. In an ideal gas, in which all interactions between the molecules can be neglected because V_m is sufficiently large, only the first term, unity, survives on the right-hand side. *See* GAS.

The equation is important because there are rigorous relations between the coefficients B_2, B_3, and so on, as well as the interactions of the molecules in pairs, triplets, and so forth. It provides a valuable route to a knowledge of the intermolecular forces. Thus if the intermolecular energy of a pair of molecules at a separation r is $u(r)$, then the second virial coefficient can be expressed as Eq. (2). Here

$$B_2(T) = 2\pi N_A \int_0^\infty (1 - e^{-u(r)/kT}) r^2\, dr \qquad (2)$$

$N_A = 6.0221 \times 10^{23}$ mol^{-1} is Avogadro's constant, and $k = R/N_A$. In a gas mixture, B_n is a polynomial of order n in the mole fractions of the components. *See* INTERMOLECULAR FORCES.

The virial equation is useful in practice because it represents the pressure accurately at low and moderate gas densities, for example, up to about 4 mol L^{-1} for nitrogen at room temperature, which corresponds to a pressure of about 100 atm (10 MPa). It is not useful at very high densities, where the series may diverge, and is inapplicable to liquids. It can be rearranged to give the ratio pV_m/RT as an expansion in powers of the pressure instead of the density, which is equally useful empirically, but the coefficients of the pressure expansion are not usually called virial coefficients, and lack any simple relation to the intermolecular forces or, in a mixture, to the composition of the gas. *See* VAN DER WAALS EQUATION. J. S. Rowlinson

Bibliography. K. E. Bett, J. S. Rowlinson, and G. Saville, *Thermodynamics for Chemical Engineers*, MIT Press, Cambridge, MA, 1975; E. A. Mason and T. H. Spurling, *The Virial Equation of State*, Pergamon, Oxford, 1969.

Virial theorem

A theorem in classical mechanics which relates the kinetic energy of a system to the virial of Clausius, as defined below. The theorem can be generalized to quantum mechanics and has widespread application. It connects the average kinetic and potential energies for systems in which the potential is a power of the radius. Since the theorem involves integral quantities such as the total kinetic energy, rather than the kinetic energies of the individual particles that may be involved, it gives valuable information on the behavior of complex systems. For example, in statistical mechanics it is intimately connected to the equipartition theorem; in astrophysics it may be used to connect the internal temperature, mass, and radius of a star and to discuss stellar stability. The virial theorem makes possible a very easy derivation of the somewhat counterintuitive result that as a star radiates energy and contracts it heats up, rather than cooling down. *See* STAR; STATISTICAL MECHANICS; STELLAR EVOLUTION.

Derivation. The derivation of the virial theorem begins with the sum, over all the particles of the system, of the dot product of the position ($\mathbf{r}$) and momentum ($\mathbf{p}$). This quantity, labeled G, is given in Eq. (1),

$$G \equiv \sum_i \mathbf{p}_i \cdot \mathbf{r}_i \qquad (1)$$

where the index i runs over all the particles in the system. (The quantity G is half of the time derivative of the spherical moment of inertia.) The time derivative of the right-hand side of Eq. (1) involves two terms. Taking the time derivatives of the momenta, and noticing that these time derivatives are the forces, leads to the last term in Eq. (2). Taking the

$$\frac{dG}{dt} = 2T + \sum_i \mathbf{F}_i \cdot \mathbf{r}_i \qquad (2)$$

time derivatives of the positions to get the velocities, and noticing that the momenta times the velocities just gives twice the kinetic energy, T, leads to the other term on the right-hand side of Eq. (2). Thus, Eq. (2) is obtained. *See* CALCULUS OF VECTORS; ENERGY; FORCE; KINETICS (CLASSICAL MECHANICS); MOMENT OF INERTIA.

To complete the derivation, this result must be averaged over time. Since the left-hand side of Eq. (2) is the time derivative of G, its average is the difference between the values of G at the beginning and end of the interval divided by the interval. If the system is periodic, it can be averaged over a period. Then the difference of the G's vanishes, and so does

the average. If the system is confined, so that G is bounded, by which is meant that the velocities and position vectors remain finite, then the system can be averaged over a long time. The average of the left-hand side then approaches 0 as the time interval approaches infinity. In either case, the virial theorem (3) is obtained, in which the angle brackets indicate

$$2\langle T\rangle = -\sum_i \langle \boldsymbol{F}_i \cdot \boldsymbol{r}_i\rangle \tag{3}$$

averages. In Eq. (3), the right-hand side is twice the virial of Clausius.

Generalizations. Forces proportional to the velocity, such as those due to a uniform external magnetic field, have no effect on the validity of the virial theorem. Indeed, the force exerted on a particle by a magnetic field is the cross product of the field and the particle's velocity; and the sum, over all the particles in the system, of the dot product of this force and the particle's position is a total derivative and averages to zero if G is either periodic or bounded, as has been assumed. In quantum mechanics, the commutator with the hamiltonian is used in place of the time derivative to obtain the same result. A relativistic derivation is somewhat more difficult. It can be done by integrating the divergence of the energy-momentum tensor dotted into a coordinate. For electromagnetic or gravitational forces, it is found that the total energy, E, is given by Eq. (4), where m_i and

$$E = \sum_i m_i c^2 \left(\frac{1 - v_i^2}{c^2}\right)^{1/2} \tag{4}$$

v_i are the mass and velocity of the ith particle and c is the velocity of light. The correct nonrelativistic result is easily obtained from this equation. *See* NONRELATIVISTIC QUANTUM THEORY; RELATIVITY.

Elementary results. In the common case that the forces are derivable from a power-law potential, V, proportional to r^k, where k is a constant, the virial is just $-k/2$ times the potential energy. Thus, in this case the virial theorem simply states that the kinetic energy is $k/2$ times the potential energy. For a system connected by Hooke's-law springs, $k = 2$, and the average kinetic and potential energies are equal. For $k = -1$, that is, for gravitational or Coulomb forces, the potential energy is minus twice the kinetic energy. *See* COULOMB'S LAW; GRAVITATION; HARMONIC MOTION.

Stars. Stars have only small electric fields because the electrical forces keep the electrons and the positive ions close together. Hence the dominant source of the potential energy is gravity. Now, in the derivation of the virial theorem, the kinetic energy is just half the sum of products of the masses and velocity squares; that is, T involves the translational kinetic energy, but not the internal kinetic energy of the molecules. Thus, it is simply related to the temperature. Other degrees of freedom, rotation and vibration of molecules, for instance, make additional contributions to the internal energy. The total internal energy may be written as in Eq. (5), recalling the usual thermodynamic definition

$$U = \frac{2T}{3(\gamma - 1)} \tag{5}$$

of γ as the ratio of the specific heats at constant pressure and constant volume. The difference between the specific heats is the gas constant, R, so that γ gets smaller as the specific heats get larger. For a nonrelativistic monatomic gas or plasma, $\gamma = 5/3$. For a relativistic gas or for photons, $\gamma = 4/3$. Endothermic reactions such as the ionization of atoms or the breakup of nuclei, both of which occur as the temperature increases, reduce the value of γ. *See* KINETIC THEORY OF MATTER; SPECIFIC HEAT.

As the star loses energy, for instance by radiation, the total energy, E, decreases. But since the total energy is related to T by Eq. (6), T may increase.

$$E = V + U = \left[-2 + \frac{2}{3(\gamma - 1)}\right] T \tag{6}$$

For γ greater than 4/3, the gravitational potential energy, V, becomes more negative; the internal energy, U, becomes more positive; and the temperature increases.

Thermal disintegration of iron nuclei sets in at several billion kelvins and drives γ below 4/3. When γ falls below 4/3, since T is certainly positive, it appears that the total energy must also be positive. This would mean that energy had to be supplied to assemble the star, which is unreasonable, since nothing does work on interstellar material to assemble it into stars. There is a fundamental error in assuming that the star is nearly static, that is to say, the time derivative of G may be neglected in Eq. (2) (or, equivalently, the second time derivative of the spherical moment of inertia may be neglected). While that is an excellent approximate for ordinary stars whose configuration evolves very slowly, it must be concluded that stars with less than 4/3 are unstable, and lose pressure support and collapse precipitously.

In real stars, γ approaches 4/3 only when the electrons become relativistic, so that the connection between kinetic energy and velocity is different from what has been used here. A more rigorous analysis, however, leads to the same conclusion. When γ falls below 4/3 in the stellar core, the core collapses and a supernova results. *See* GRAVITATIONAL COLLAPSE; SUPERNOVA.

Albert G. Petschek

Bibliography. D. D. Clayton, *Principles of Stellar Evolution and Nucleosynthesis*, 1983; G. W. Collins, II, *The Virial Theorem in Stellar Astrophysics*, 1978; H. Goldstein, *Classical Mechanics*, 3d ed., 2001; R. K. Pathria, *Statistical Mechanics*, 2d ed., 1996.

Viroids

Small infectious nucleic acid particles containing independently replicating single-stranded circular ribonucleic acids (RNAs) of 246–399 nucleotides, which are known to infect certain monocot and dicot plants. Viroids are the smallest known disease agents, with an estimated molecular weight (sum

of the atomic weights of a molecule's constituent atoms) of as little as $1.0–1.3 \times 10^4$, in marked contrast to the conventional plant virus genomes, which have molecular weights of approximately 2×10^6. Conventional viruses are made up of nucleic acid encapsulated in protein (capsid), whereas viroids are uniquely characterized by the absence of a capsid. In spite of their small size, viroid RNAs can replicate and produce characteristic disease syndromes when introduced into cells. Although the viroids thus far identified are associated with plants, based on the unusual properties of the infectious agents they may also be found to affect other forms of life, including humans. It is possible that in instances where a viral etiology of disease has been assumed but where no causal virus has been identified, viroid infectious agents are involved.

Classification. Up to 27 different viroid species have been described from widely separated geographical locations. Viroids are classified in two families. The first family, Pospiviroidae, contains 24 viroids which have a central conserved nucleotide region which contain the same nucleotide sequences. Some examples include potato spindle tuber viroid (PSTV), citrus exocortis viroid (CEV), chrysanthemum stunt viroid (CSV), and cucumber pale fruit viroid (CPFV). The second family, Avsunviroidae, contains three viroids, which do not have a central conserved region: avocado sunblotch viroid (ASV), peach latent mosaic viroid (PLMV), and chrysanthemum chlorotic mottle viroid (ChCMV).

Diseases and hosts. Diseases caused by viroids are agriculturally relevant, affecting economically important plants and an assortment of herbaceous and woody plants, such as potato, tomato, cucumber, hop, citrus, grapevine, coconut palm, fruit trees (apple, pear, plum, peach, avocado), and ornamentals (chrysanthemum and coleus).

Viroid infections in some plant species produce profound disease symptoms ranging from stunting and leaf epinasty (downward bending) to plant death, whereas infections in other species produce few detectable symptoms compared with uninoculated control plants. Viroids generally have a restricted host range, although several viroids can infect the same hosts and cause similar symptoms. For example, PSTV, CEV, CSV, and CPFV infect Rutgers tomato (*Lycopersicon esculentum*), and all except CSV cause stunting of growth, epinasty, and bunchy top. Infection with PSTV, CEV, and CSV produces similar symptoms in purple passion plant (*Gynura aurantiaca*).

Good controls are not available for diseases caused by these small infectious agents other than indexing procedures to provide viroid-free propagules. Indexing programs generally rely on the use of bioassays that are laborious and require incubation periods of several months. Sensitive biochemical assays [such as polyacrylamide gel electrophoresis, dot-blot hybridization with labeled complementary deoxyribonucleic acid (cDNA) or complementary RNA (cRNA) probes, and reverse transcription polymerase chain reaction] are faster and more reliable to use than conventional bioassays and have been developed to detect infectious viroid RNAs in infected host tissue.

Cross-protection. Although viroids have been shown to differ significantly in their nucleotide sequence (CCR of Pospirviroidae is the only portion of the viroid molecule which has conserved nuclear sequences), the phenomenon of cross-protection is known to occur in some viroid double infections. Cross-protection is defined here as the interference in symptom expression of a challenge-inoculated viroid or virus in a previously infected plant. Cross-protection has been shown to exist in viroid species of both Pospiviroidae and Avsunviroidae families and is known to be due to subtle variations in the structure of their RNA. As a result of these phenomena, plants infected with a latent or nonsymptomatic strain become protected against the challenge inoculation by a severe strain. The practical applications of cross-protection have been to indicate relationships among plant viruses and to control tobacco mosaic virus in commercial tomato plantings and tristeza of citrus in Brazil.

Cross-protection has been shown between strains of PSTV and also between strains of ChCMV. Moreover, cross-protection has been shown between strains of the closely related viroids PSTV, CEV and CSV. However, a lack of cross-protection has been shown for ChCMV strains against PSTV, CEV and CSV strains, which separates the former from the latter three viroid species (two different viroid families).

Interference with other pathogens. Interference has also been found between viroids and bacteria in double infections in chrysanthemums. The amount of pith maceration of chrysanthemum cuttings, cultivar Bonnie Jean, by *Erwinia chrysanthemi* 159 is significantly reduced by prior CSV infection. The reduction is apparent in CSV-infected plants approximately 30 days after inoculation with the viroid. This period of time coincides with CSV symptom expression and represents a period of viroid multiplication. It appears that a threshold of viroid molecules within infected tissue is necessary for three observed events: symptom development, viroid detection by polyacrylamide gel electrophoresis assay, and decreased bacterial pith maceration. This reaction differs from the cross-protection phenomenon, which involves closely related organisms (that is, strains of a single virus and viroid). Furthermore, reduced infection has been found with *Phytophthora infestans* (which causes potato late blight) in potato plants infected with PSTV. Viruses, however, have been reported to reduce or increase the severity of fungal infections. *See* PLANT VIRUSES AND VIROIDS; VIRUS.

R. K. Horst

Bibliography. T. O. Diener (ed.), *The Viroids*, 1987; T. O. Diener, *Virus Genes*, 11:119–131, 1996; R. Flores et al. (eds.), *Advances in Virus Research*, vol. 55, 2001; T. C. Hall and J. W. Davies (eds.), *Nucleic Acids in Plants*, vol. 2, 1979; J. S. Semancik et al. (eds.), *Viroids and Viroid-like Pathogens*, 1988.

Virtual acoustics

The simulation of the complex acoustic field experienced by a listener within an environment. The technology is also known as three-dimensional sound and auralization. Going beyond the simple left-right volume adjustment of normal stereo techniques, the goal is to process sounds so that they appear to come from particular locations in three-dimensional space. Although loudspeaker systems have been developed, much of the work in the field focuses on using headphones for playback and is the outgrowth of earlier analog techniques. For example, in binaural recording, the sound of an orchestra playing classical music is recorded through small microphones in the two imitation ear canals of an anthropomorphic artificial or dummy head placed in the audience of a concert hall. When the recorded piece is played back over headphones, the listener passively experiences the illusion of hearing the violins on the left and the cellos on the right, along with all the associated echoes, resonances, and ambience of the original environment. Techniques use digital signal processing to synthesize the acoustical properties that people use to localize a sound source in space. Thus, they provide the flexibility of a kind of digital dummy head, allowing a more active experience in which a listener can both design and move around or interact with a simulated acoustic environment in real time. *See* BINAURAL SOUND SYSTEM.

Psychoacoustical cues. The success of virtual acoustics is critically dependent on whether the acoustical cues used by humans to locate sounds have been adequately synthesized. Much of the present understanding of sound localization is based on Lord Rayleigh's duplex theory of 1907, which describes the role of two primary cues: interaural time differences for low-frequency signals, and interaural intensity differences for high-frequency signals. However, the duplex theory neglects the important interaction between incoming sound waves and the outer ears, or pinnae: spectral shaping by the pinnae is highly direction dependent; the absence of pinna cues degrades localization accuracy; and pinna cues are at least partially responsible for externalization, or the outside-the-head sensation. *See* EAR (VERTEBRATE).

There may be many cumulative effects on the sound as it makes its way to the eardrum, but all of these effects can be expressed as a single filtering operation much like the effects of a graphic equalizer in a stereo system. The exact nature of this filter can be measured by an experiment in which an impulse (a single, very short sound pulse or click) is produced by a loudspeaker at a particular location. The acoustic shaping by the two ears is then measured by recording the outputs of small probe microphones placed inside the ear canals of an individual or an artificial head (**Fig. 1**). If the measurement of the two ears occurs simultaneously, the responses, when taken together as a pair of filters, include an estimate of the interaural differences as well. Thus, this technique makes it possible to measure all of the relevant spatial cues together for a given source location, for a given listener, and in a given room or environment (**Fig. 2**). *See* ELECTRIC FILTER; EQUALIZER.

The spatial filters constructed from these ear-dependent characteristics are examples of finite impulse response filters, and the characteristics themselves are often referred to as head-related transfer functions (HRTFs). Filtering in the frequency domain is a point-by-point multiplication of the input signal with the left and right head-related transfer

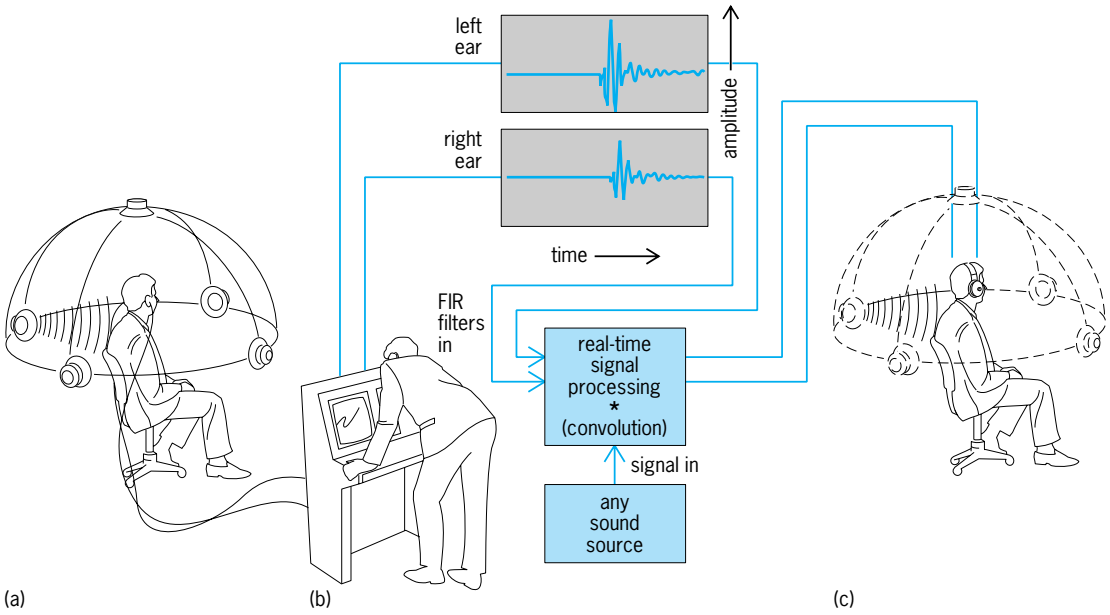


Fig. 1. Technique for synthesizing virtual acoustic sources with measurements of the head-related transfer function. (*a*) Pinnae (outer-ear) responses are measured with probe microphones. Placement of loudspeakers is illustrative only. (*b*) Pinnae transforms are digitized as finite-impulse-response (FIR) filters. An example of a pair of finite impulse responses measured for a source location at 90° to the left and 0° elevation (at ear level) is shown for the left and right ears. (*c*) Synthesized cues are played over headphones. (*After E. M. Wenzel, Localization in virtual acoustic displays, Presence: Teleoper. Virt. Environ., 1(1):80–107, 1992*)

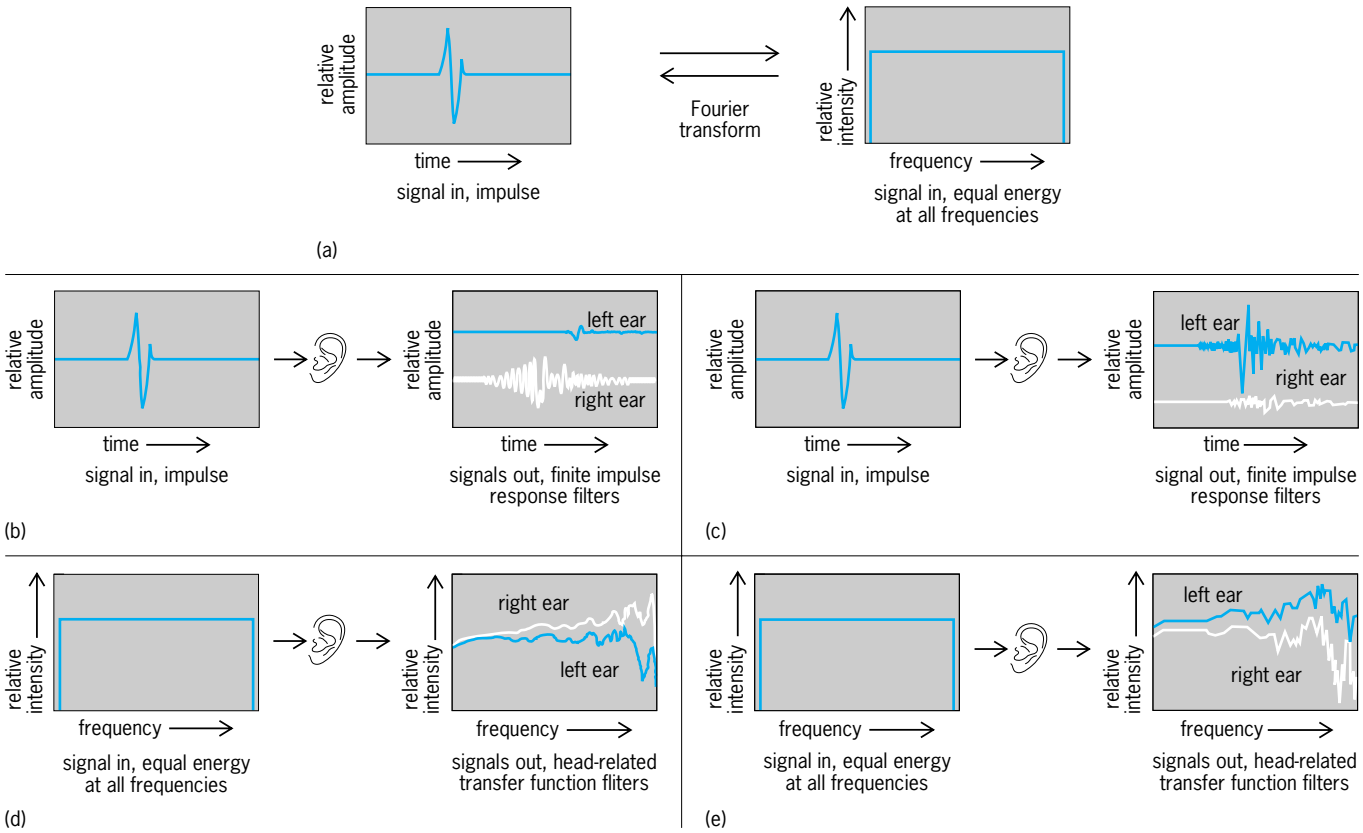


Fig. 2. Effects of spectral shaping by the pinnae (outer-ear structures). (*a*) Equivalent representations, via the Fourier transform, of a broad-band acoustic signal before interaction with the outer-ear structures, in the time domain (where the acoustic signal is an impulse) and the frequency domain (where only intensity is shown here). (*b*) Effect of the pinnae on an impulse delivered from a loudspeaker directly to the right (+90° azimuth, 0° elevation), as measured in the left-ear and right-ear canals of the individual. (*c*) Effect of the pinnae on an impulse delivered from a loudspeaker 60° to the left (−60° azimuth, 0° elevation). (*d*) Frequency-domain representation of the effect on the signal from 90° azimuth, 0° elevation. (*e*) Frequency-domain representation of the effect on the signal from −60° azimuth, 0° elevation. (*After E. M. Wenzel, Localization in virtual acoustic displays, Presence: Teleoper. Virt. Environ., 1(1):80–107, 1992*)

functions, while filtering in the time domain by the finite impulse response filters occurs via a more complex multiply-and-add operation known as convolution. *See* CONTROL SYSTEMS; FOURIER SERIES AND TRANSFORMS; INTEGRAL TRANSFORM.

By using these spatial filters, it should be possible to impose spatial characteristics on a signal so that the signal apparently emanates from the originally measured location. Of course, the localizability of the sound will also depend on other factors such as its original spectral content. Narrowband sounds (such as sine waves) are generally hard to localize, while broadband, impulsive sounds are the easiest to locate. Other acoustic features, such as the ratio of direct to reflected energy in a reverberant field, can also provide a cue to distance (closer sources correspond to larger ratios) as well as enhance the sensation of externalization. *See* REVERBERATION.

Psychophysical validation. The only conclusive test of the simulation technique is an operational one in which the localization of real and synthesized sources is directly compared. Two kinds of errors are usually observed when subjects are asked to judge the position of a stationary sound source in the free field: a relatively small error in resolution on the order of about 5–20°, and front-back reversals, in which a source in the front (or rear) is perceived by the listener as if it were in the rear (or front). Occasionally, reversals in elevation are also observed.

In studies that compare localization judgments of stationary broadband sources in the free field with judgments of virtual (headphone-presented) sources synthesized from the subjects' own head-related transfer functions, localization accuracy is generally similar for the free-field and headphone stimuli, although source elevation is less well defined and reversal rates increase from about 5% for real sources to about 10% for virtual sources. Other experiments suggest that most listeners can obtain useful directional information from nonpersonalized head-related transfer functions, particularly for the left-right dimension of azimuth, so long as these functions come from a "good localizer," who accurately localizes both real and virtual sources. Again, localization is comparable for free-field and headphone stimuli, although, for some subjects, source elevation is completely disrupted and reversal rates increase substantially. Interestingly, some people show little ability to localize source elevation for both real and virtual sounds, as if they have poorly designed pinna structures which have little impact on the spectra of incoming sounds. Listening "through" these

people's transfer functions results in a loss of elevation perception for any listener.

Real-time implementation. Real-time systems began to appear only when powerful digital signal processing chips were available. In general, these systems are intended for headphone presentation and use time-domain convolution to achieve real-time performance. Usually, one or more sources are simulated for an anechoic environment by using spatial filters chosen according to the output of a head-tracking device in order to compensate for the listener's position and orientation. Smooth motion trajectories can be simulated by interpolation between impulse responses originally measured at intervals of about 10° or more, and a simple distance cue may be provided via real-time scaling of amplitude. While such systems simulate only the direct paths from each virtual source to the listener, it is possible to freely manipulate both source and listener position in a highly dynamic, interactive display. *See* INTEGRATED CIRCUITS.

The same degree of interactivity is desirable in a more complex simulation since perceptual research suggests that errors like front-back reversals and failures of externalization can be mitigated by providing the acoustic cues available in reverberant environments. Of particular interest is the development of the image model and related ray-tracing techniques for simulating room characteristics. Extensions of these techniques implement the directional characteristics of source reflections in real time by convolution with filters based on head-related transfer functions. For example, in one implementation of the image model the walls, floor, and ceiling in an environment are simulated by calculating the mirror image of a sound source behind each surface to account for the specular reflection of the source signal. The filtering effect of surfaces such as wood or drapery can also be modeled with a separate filter whose output is delayed by the time required for the sound to propagate from each image source being represented. Finally, the delayed signals are processed with spatial filters appropriate to the direction of each reflection to create the components of the binaural output. The superposition of these components with the direct path to the source is then presented over headphones. Such dynamic modeling requires enormous computational resources for real-time implementation in a truly interactive (head-tracked) display. *See* ARCHITECTURAL ACOUSTICS; HEARING (HUMAN); PSYCHOACOUSTICS; REFLECTION OF SOUND; SOUND; TELECONFERENCING; VIRTUAL REALITY.

Elizabeth M. Wenzel

Bibliography. J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, 1983; F. L. Wightman and D. J. Kistler, Headphone simulation of free-field listening, I: Stimulus synthesis, *J. Acous. Soc. Amer.*, 85(2):858–867, 1989; F. L. Wightman and D. J. Kistler, Resolution of front-back ambiguity in spatial hearing by listener and source movement, *J. Acous. Soc. Amer.*, 105(5):2841–2583, 1999.

Virtual manufacturing

The modeling of manufacturing systems using audiovisual or other sensory features to simulate or design alternatives for an actual manufacturing environment, or the prototyping and manufacture of a proposed product using computers. The motivation for virtual manufacturing is to enhance people's ability to predict potential problems and inefficiencies in product functionality and manufacturability before real manufacturing occurs.

The term "virtual manufacturing" became prominent in the early 1990s, in part because of the U.S. Department of Defense Virtual Manufacturing Initiative. The concept has since gained international acceptance and has broadened in scope. For the first half of the 1990s, pioneering work in this field was done by a handful of major organizations, mainly in the aerospace, earthmoving equipment, and automobile industries, plus a few specialized academic research groups. Factors accelerating the worldwide market interest in virtual manufacturing include price-performance improvements in the hardware and software technologies required, and increased awareness of its huge potential. *See* MANUFACTURING ENGINEERING; MODEL THEORY.

Framework. The concepts underlying virtual manufacturing include virtual reality, high-speed networking and software interfaces, agile manufacturing, and rapid prototyping.

Virtual reality. Virtual reality is broadly defined as the ability to create and interact in cyberspace, that is, a simulated space that represents an environment very similar to the actual environment. The subset of virtual reality that is used in virtual manufacturing is commonly known as virtual environment. Virtual environment systems differ from previously developed computer-centered systems in the extent to which real-time interaction is simulated. The perceived visual space is three-dimensional rather than two-dimensional, the human-machine interface is multimodal, and the user is immersed in the computer-generated environment; the screen separating the user and the computer becomes invisible to the user.

The virtual environment for virtual manufacturing is simulated through immersion in computer graphics coupled with an acoustic interface, domain-independent interacting devices such as wands, and domain-specific devices such as steering and brakes for cars or earthmovers or instrument clusters for airplanes. There are two main techniques for creating immersion in current virtual reality systems: head-mounted displays and stereoscopic projectors. A head-mounted display is a tracking device incorporating liquid-crystal displays or miniature cathode-ray tubes which is mounted on the head of the user to provide information on head movements for updating the visual images. An example is the Binocular Omni-Oriented Monitor (BOOM) developed commercially. Stereoscopic projectors provide a three-dimensional sensation using screen

projections of two distinct views of an object separated by a small angle based on viewing the object with the left eye only and with the right eye only. The users normally wear liquid-crystal-display shutter glasses. An example is the Cave Automatic Virtual Environment (CAVE) developed by the University of Illinois at Chicago. *See* VIRTUAL REALITY.

High-speed networking and software interfaces. The issues here are concerned with computer-aided-design (CAD) model portability among systems, trade-offs of high-detail models versus real-time interaction and display, rapid prototyping, collaborative design using virtual reality over distance, use of the Web for small- or medium-business virtual manufacturing, use of qualitative information (illumination, sound levels, ease of supervision, handicap accessibility) to design manufacturing systems, use of intelligent and autonomous agents in virtual environments, and the validity of virtual reality versus reality (quantitative testing of virtual versus real assemblies or equipment). With regard to hardware, the two extremes (in terms of cost and performance) are personal computers using Virtual Reality Modeling Language (VRML) and the Internet, and workstations using the CAVE library and asynchronous transfer mode (ATM) networks with data transfer speeds up to 622 Mbps (see **illus.**).

Agile manufacturing. A term that is sometimes mentioned in the context of virtual manufacturing is agile manufacturing. Agile manufacturing integrates an organization's people and technologies through innovative management and organization, knowledgeable and empowered people, and flexible and intelligent technologies. Virtual manufacturing provides a model for making rapid changes in products and processes based on customer requirements, and an agile manufacturing system attempts to implement it.

Rapid prototyping. An area in which virtual manufacturing has made an impact is in rapid prototyping, or layered manufacturing. In processes such as stereolithography, selective laser sintering, and fused deposition modeling, a CAD drawing of a part is processed to create a layered file of the part. The part is built one layer at a time, precisely depositing layer upon layer of material.

Virtual collaborative environments. Global virtual manufacturing extends the definition of virtual manufacturing to include the use of the Internet and intranets (global communications networks) for virtual component sourcing, and the use of virtual collaborative design and testing environments by multiple organizations or sites. Companies committed to global virtual manufacturing may shorten the time to market for new products, cut the cost of prototyping and preproduction engineering, enable many more variations to be tried out before committing to manufacture, and increase the range and effectiveness of quality assurance testing. Prototypes can be virtually assembled, tested, and inspected as part of production planning and operative training procedures. They can be demonstrated, market-tested, used to brief and train sales and customer staff, transmitted instantly from site to site via communications links, and modified and recycled rapidly in response to feedback.

Demonstration of Virtual Reality Modeling Language (VRML) Browser interface with CAVE™ for 3D Immersive Analysis of Manufacturing Shopfloor. The CAVE™ virtual reality display features three room-sized rear-projected walls and a top-projected floor, each driven by workstations. (CAVE is a trademark of the University of Illinois.)

Manufacturers and their worldwide subcontractors and main suppliers can establish agile manufacturing teams that will work together on the design, virtual prototyping, and simulated assembly of a particular product, at the same time establishing confidence in the virtual supply chain. Using the most advanced virtual reality systems, geographically remote members of the team can meet in the same virtual design environment to discuss and implement changes to virtual prototypes.

An example of a recent development in virtual collaborative environments is the projection of gestures and movements of multiple remote designers through voice-activated avatars to help explain the designers' intentions to others in real-time using high-speed ATM networks.

Monitoring and control for complex manufacturing. For monitoring and control of complex manufacturing systems, four dimensions can be conceived to express complexity: space, time, process, and network. Space concerns the physical location, layout, and flow issues critical in all manufacturing operations. Time involves the facility life cycle and operational dynamics beginning with the concurrent engineering of the production process and testing facilities during product design, extending through the maturation and decline of the first-generation products, and cycling through the same process for future generation products. Process relates to the coherent integration of engineering, management, and manufacturing processes. It permits the examination of the important yet intricate interplay of relationships between classically isolated functions,

such as the relationships between production planning and purchasing, production control and marketing, quality and maintenance, and design and manufacturing. Process decisions range from long-range operational planning to short-term machine- and device-level planning and control. The integration between various levels of aggregation is essential. Network concerns organization and infrastructure integration. While the third dimension, process, focuses on the actions, this dimension concentrates on the actors and their needs and responsibilities. Clearly including personnel, the set of actors also includes all devices, equipment, and workstations; all organizational units, be they cells, teams, departments, or factories; and all external interactors, such as customers, vendors, subcontractors, and partners. Issues such as contrasting hierarchically controlled networks with hierarchical, autonomous agent networks must be addressed. *See* CONCURRENT PROCESSING.

Virtual manufacturing techniques address issues such as designing products that can be evaluated and tested for structural properties, ergonomic functionality, and reliability without having to build actual scale models; designing products for esthetic value meeting individual customer preference; ensuring facility and equipment compliance with various federally mandated standards; facilitating remote operation and control of equipment (telemanufacturing and telerobotics); developing processes to ensure manufacturability without having to manufacture the product (for example, avoiding destructive testing); developing production plans and schedules and simulating their correctness; and educating employees on advanced manufacturing techniques with emphasis on safety.

Applications. Virtual manufacturing has been used in the automotive industry to design the interior of cars and perform virtual crash testing. The heavy-equipment industry has built electronic prototypes of earthmoving equipment instead of physical prototypes. Both industries are actively exploring virtual collaborative environments using multiple CAVEs to coordinate designs with their suppliers and showrooms. The pharmaceutical industry is exploring use of this technology for drug manufacturing to rapidly conform to the Federal Drug Administration (FDA) requirements. The electronics industry is comparing the impact of employees trained with head-mounted devices in virtual reality environments with those trained in the real environment. The telecommunications industry is using haptic interfaces, which allow users to feel their interfaces and interactions, to design their phones.

Pat Banerjee

Bibliography. P. Banerjee, L. Freitag, and S. Mehrotra (eds.), *Virtual Reality in Manufacturing Research and Education: Proceedings of the Symposium Sponsored by the National Science Foundation*, 1996; A. Johnson et al., CAVERN: The CAVE Research Network, *Proceedings of the 1st International Symposium on Multimedia Virtual Laboratory*, Tokyo, March 25, 1998; Special Issue on Virtual Manufacturing, *IIE Trans.*, 30(7):581–644, 1998.

Virtual reality

A form of human-computer interaction in which a real or imaginary environment is simulated and users interact with and manipulate that world. Users travel within the simulated world by moving toward where they want to be, and interact with things in the world by grasping and manipulating simulated objects. In the most successful virtual environments, users feel that they are truly present in the simulated world and that their experience in the virtual world matches what they would experience in the environment being simulated. This sensation is referred to as engagement, immersion, or presence, and it is this quality that distinguishes virtual reality from other forms of human-computer interaction. *See* HUMAN-COMPUTER INTERACTION.

When a user interacts with a virtual environment, the computer-generated graphics display must be updated with each turn of the head or movement of the hand. The virtual environment must be able to generate and display realistic-looking views of the simulated world quickly enough that the interaction feels responsive and natural. Initially, use of the technology was limited to those applications where the cost of high-performance computer graphics hardware could be justified, such as in flight simulators. The decreasing cost of such hardware has made it possible to experiment with applying the technology in many other areas. *See* COMPUTER GRAPHICS.

Hardware. Virtual reality relies on a variety of specialized input and output devices to achieve a sense of natural interaction in a simulated environment. The goal is to go beyond traditional computer interfaces and mimic human experience with the real world. For example, in traditional computer-generated graphics, only one view of a scene is projected, rendered, and drawn on the display screen. But when a person views the real world, the left and right eyes see different views and the perceptual system uses these differences to judge the positions of objects in the scene. Also, human interaction with the world is quite rich; people move around, reach for things, pick them up, and manipulate objects directly. All the senses are involved. But in traditional computer applications, user interaction modes are limited to typing on a keyboard, moving a mouse around on a flat surface, or perhaps using a handheld joystick. A primary goal of virtual reality is to develop strategies for creating a natural interaction environment that will contribute to a sense of presence in the simulated world. *See* STEREOSCOPY.

Tracker. The most important of the input devices used in a virtual environment, a tracker is capable of reporting its location in space and its orientation. Tracking devices can be optical, magnetic, or acoustic. Optical systems use cameras to record user activity. The images are interpreted to detect body position or hand gestures. Camera-based environments allow the user to direct system activity in a very natural and unencumbered manner, but there are a number of difficult problems with this approach. For

example, it is difficult to interpret images of hand activity when users move their hands quickly or when fingers overlap. *See* COMPUTER VISION.

Acoustic and magnetic tracking systems use a stationary element and small movable elements. The movable element is usually worn on the user's head or hand and is connected by cable to the computer system. In acoustic systems, the movable tracker element emits a high-frequency sound, which is detected by a set of stationary microphones. The strength of the signal recorded by each microphone will depend on the distance between the sound source and a particular microphone, and the variation in signal strength indicates the location of the sound. Acoustic systems are accurate to within a fraction of an inch (a few millimeters). In magnetic tracking systems, the stationary element emits a pulsed magnetic field. The movable sensor attached to the user reports its position and orientation relative to the source. Magnetic systems are accurate to 0.1 in. (2.5 mm) in position and 0.1° in rotation. Optical and acoustic systems rely on a clear line-of-sight path between the signal source and the receptor. Magnetic tracking systems do not have this requirement, but magnetic tracking can be susceptible to interference from metallic objects in the environment. *See* MICROPHONE.

A tracker is sometimes combined with a traditional computer input device, such as a mouse or a joystick. The user holds the device and can use the regular controls, such as mouse buttons, to signify when different actions should occur. Everyday objects, including treadmills and stationary bicycles, have also been adapted to serve as input devices to virtual reality systems. *See* COMPUTER PERIPHERAL DEVICES.

Data glove. An attempt to provide a truly natural input device, the data glove is outfitted with sensors that can read the angle of each of the finger joints in the hand. Wearing such a glove, users can interact with the virtual world through hand gestures, such as pointing or making a fist. Gages that measure strain or fiber optics, where the amount of transmitted light signifies the angle at a finger joint, are used in gloves to determine finger positions. *See* FIBER-OPTIC SENSOR; STRAIN GAGE.

Stereoscopic displays. The real-world visual experience is approximated in virtual environments by using stereoscopic displays. Two views of the simulated world are generated, one for each eye, and a stereoscopic display device is used to show the correct view to each eye. Two types of stereo viewing devices are the head-mounted display and the binocular omni-orientation monitor (BOOM). Each device has two display screens, and the appropriate image is shown to the left and right eye.

Stereo glasses, lightweight and inexpensive, provide another alternative. High-quality stereo glasses usually use liquid-crystal shutters that allow light to pass through only one lens at a time. Computer-generated views for the left and right eyes are alternated very quickly on a computer monitor, or perhaps on a large projection screen. A transmitter connected to the computer synchronizes the glasses to the display so that the appropriate lens is active at each instant. Switching between views is fast enough that the user sees the overall scene in stereo and is unaware of the alternating views. *See* LIQUID CRYSTALS.

The different display devices have advantages and disadvantages. The BOOM user does not need to wear any special equipment, which allows the user to look away from the display at any time. It also makes it easy for multiple users to take turns at the BOOM. However, the ability to step out of the virtual world at any moment can decrease the feeling of presence or immersion. Head-mounted displays, however, are helmetlike devices that are strapped on and require some effort to remove. Lightweight stereo glasses are easy to wear and to remove. Compared to the other devices, glasses are inexpensive enough that it is reasonable to provide a pair for everyone in a group.

Applications. Virtual reality can be applied in a variety of ways, including entertainment, education and training, manufacturing and design, and exploring scientific data and processes. In scientific and engineering research, virtual environments are used to visually explore whatever physical world phenomenon is under study. For example, a researcher can study how air travels near the surface of a wing. Virtual environments have also been used to explore whether different molecules will bond, with a force-feedback device used to convey whether molecules are attracted or repulsed. These virtual environments simulate physical phenomena occurring in the everyday world, but virtual environments are also used to explore worlds that can only be imagined, such as four-dimensional geometries, or microworlds that operate according to their own set of physical laws.

Training personnel for work in dangerous environments or with expensive equipment is best done through simulation. Airplane pilots, for example, train in flight simulators. Virtual reality can enable medical personnel to practice new surgical procedures on simulated individuals. As a form of entertainment, virtual reality is a highly engaging way to experience imaginary worlds and to play games. *See* AIRCRAFT TESTING; VIDEO GAMES.

Virtual reality also provides a way to experiment with prototype designs for new products. For example, the traditional design process for a new car involves building physical prototypes to evaluate the design. Virtual reality provides a way to simulate the new design and experiment with it, avoiding the cost of building the physical prototype. *See* COMPUTER-AIDED DESIGN AND MANUFACTURING.

M. Pauline Baker

Bibliography. W. Barfield and T. Furness (eds.), *Virtual Environments and Advanced Interface Design*, 1995; R. Earnshaw, M. Gigante, and J. Jones (eds.), *Virtual Reality Systems*, 1993; H. Rheingold, *Virtual Reality*, 1991; Special issue on virtual reality, *IEEE Comput. Graphics Appl.*, vol. 14, no. 1, January 1994; A. Wexelblat, *Virtual Reality Applications and Explorations*, 1993.

Virtual work principle

The principle stating that the total virtual work done by all the forces acting on a system in static equilibrium is zero for a set of infinitesimal virtual displacements from equilibrium. The infinitesimal displacements are called virtual because they need not be obtained by a displacement that actually occurs in the system. The virtual work is the work done by the virtual displacements, which can be arbitrary, provided they are consistent with the constraints of the system. *See* CONSTRAINT.

The principle of virtual work is equivalent to the conditions for static equilibrium of a rigid body expressed in terms of the total forces and torques. That is, the principle of virtual work can be derived from these conditions, and conversely. *See* EQUILIBRIUM OF FORCES; STATICS.

One advantage of the principle of virtual work is that it can serve as a basis for all of statics. In the solution of problems the principle of virtual work is often useful for eliminating the need for consideration of the forces of constraint, since these forces often are perpendicular to the virtual displacements and consequently do no work.

The following problem provides an application of the principle of virtual work (see **illus.**). A uniform plank of weight W rests on a smooth floor and leans against a smooth wall. The angle between the plank and the floor is ϕ. A weightless, inextensible string connects the lower end of the plank to the wall. The value of the tension T in the string is desired. Let $2b$ be the length of the plank. Then, as can be seen from the sketch, Eqs. (1) hold. When ϕ increases by

$$y = b\sin\phi \qquad x = 2b\cos\phi \tag{1}$$

$d\phi$, the resulting changes dx and dy in x and y, respectively, are given by Eqs. (2). According to the

$$dy = b\cos\phi\, d\phi \qquad dx = -2b\sin\phi\, d\phi \tag{2}$$

principle of virtual work, the total work done in the virtual displacement produced by $d\phi$ must be zero. The forces of reaction N_1 and N_2 are normal to the wall, since the wall is smooth. Consequently they do no work. The total work, therefore, is the work done by the forces W and T moving through their respective virtual displacements dy and dx. This work is given by Eq. (3). Solving for the tension T gives Eq. (4). The same result is obtained from the equilibrium conditions expressed in terms of forces and torques. When the principle of virtual work is used, the forces of reaction need not be considered.

$$0 = -2bT\sin\phi\, d\phi + Wb\cos\phi\, d\phi \tag{3}$$

$$T = (W/2)\cot\phi \tag{4}$$

Paul W. Schmidt

Bibliography. N. G. Chetave, *Theoretical Mechanics*, 1990; H. Goldstein, *Classical Mechanics*, 3d ed., 2001.

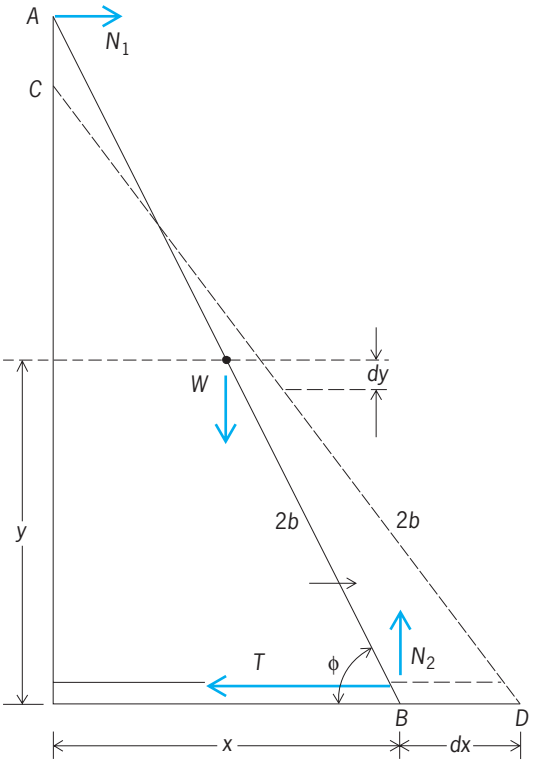


Illustration of principle of virtual work. (*After R. A. Becker, Introduction to Theoretical Mechanics, McGraw-Hill, 1954*)

Virulence

The ability of a microorganism to cause disease. Virulence and pathogenicity are often used interchangeably, but virulence may also be used to indicate the degree of pathogenicity. Infectious diseases are caused by bacteria, fungi, protozoa, viruses, and larger parasites. Scientific understanding of the underlying mechanisms of virulence has increased rapidly due to the application of the techniques of biochemistry, genetics, molecular biology, and immunology. Bacterial virulence is better understood than that of other infectious agents.

Virulence is often multifactorial, involving a complex interplay between the parasite and the host. Various host factors, including age, sex, nutritional status, genetic constitution, and the status of the immune system, affect the outcome of the parasite-host interaction. Hosts with depressed immune systems, such as transplant and cancer patients, are susceptible to microorganisms not normally pathogenic in healthy hosts. Such microorganisms are referred to as opportunistic pathogens. The attribute of virulence is present in only a small portion of the total population of microorganisms, most of which are harmless or even beneficial to humans and other animals. *See* OPPORTUNISTIC INFECTIONS.

The spread of an infectious disease usually involves the adherence of the invading pathogen to a body surface. Next, the pathogen multiplies in host tissues, resisting or evading various nonspecific host defense systems. Actual disease symptoms are from damage to host tissues caused either directly

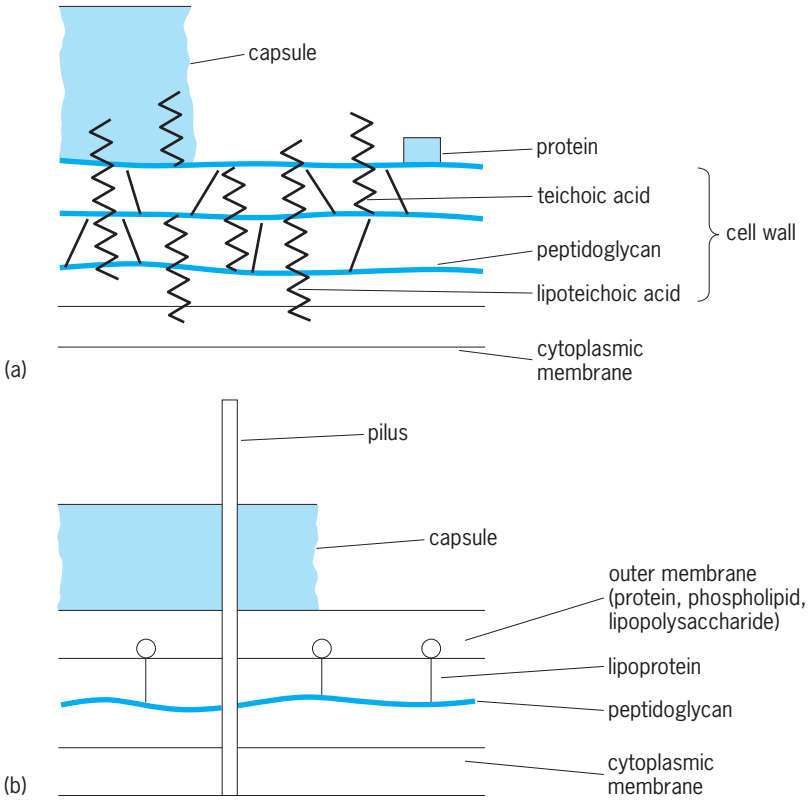
or indirectly by the microorganism's components or products.

Most genetic information in bacteria is carried in the chromosome. However, genetic information is also carried on plasmids, which are independently replicating structures much smaller than the chromosome. Plasmids may provide bacteria with additional virulence-related capabilities (such as pilus formation, iron transport systems, toxin production, and antibiotic resistance). In some bacteria, several virulence determinants are regulated by a single genetic locus. *See* PLASMID.

Microbial cell surfaces. The microbial cell surface plays a major role in virulence. It mediates adherence and resistance to host defense systems, and contains components responsible for important biological properties, including toxic ones. In both gram-positive and gram-negative bacteria a cell wall overlays the cell membrane, and in some strains a third layer known as the capsule may be present (see **illus.**) The capsule is important in allowing the microorganisms to resist the host's phagocytic defense systems. Peptidoglycan is a polymer of amino acids and amino sugars that is common to both types of bacteria, and is responsible for the rigidity of cell walls. In gram-negative bacteria a second, outer membrane is present, of which lipopolysaccharide is a characteristic component. Both peptidoglycan and lipopolysaccharide have toxic properties. Rod-like protein structures known as pili or fimbriae are present on the surface of some gram-negative bacteria that mediate adherence to host tissues. *See* BACTERIA.

Adherence. Most infections start at a mucous membrane, where pili are important virulence factors. Piliated strains of *Neisseria gonorrhoeae* are more virulent than nonpiliated strains, since pili increase attachment of the organism to human buccal and epithelial cells. In piglets a vaccine made by using pili provided effective protection against enteropathogenic *Escherichia coli* strains. In this case, antibodies to pili were responsible for resistance to infection. The interaction between host-cell receptors and microbial surface adhesins or ligands plays a marked role in the preference of certain microorganism for particular tissues, a phenomenon known as organ or tissue tropism. In a variety of bacterial pathogens, adhesion is followed by invasion of and multiplication in host cells. *See* CELLULAR ADHESION.

Nutrition. The nutritional environment provided by host tissues is an important factor in the multiplication of pathogenic microorganisms in living organisms, but is a poorly understood aspect of the host-parasite interaction. However, something is known concerning the role of iron availability in humans, and the bacterial response to it. The amount of free iron in the body is very low because various proteins in body fluids bind iron avidly. Iron is an essential bacterial nutrient, but pathogenic bacteria are nonetheless able to proliferate in this iron-poor environment. Under these conditions, bacteria often secrete low-molecular-weight molecules, siderophores, which bind iron with great affinity. Receptor proteins for



Cell surface of (*a*) gram-positive and (*b*) gram-negative bacteria.

the siderophore-iron complexes are induced in the outer membrane of gram-negative bacteria. This iron uptake system has been extensively studied in *E. coli*, and is present in a variety of other pathogenic bacteria. Commonly, the genetic information for iron uptake systems is carried on plasmids. *See* SIDEROPHORES.

Nonspecific host defenses. Two interrelated and important host defense systems are a cascade of serum proteins known as the complement system, and phagocytic cells. Host complement proteins are activated by microbial components such as lipopolysaccharide and peptidoglycan. Complement proteins promote inflammation and chemotaxis of phagocytes toward the invading organism, and render microorganisms susceptible to attachment and ingestion by phagocytes. Some complement proteins cause direct killing of gram-negative bacteria.

A variety of bacterial cell-surface components have been linked to resistance to complement-mediated killing, thereby increasing bacterial virulence. In *E. coli* these include outer membrane proteins, capsular polysaccharide, and complete lipopolysaccharide. The nature of the lipopolysaccharide is important in the resistance to complement-mediated killing of *Salmonella* species and *Shigella sonei*, while capsular polysaccharide is important in the resistance of type b *Haemophilus influenzae* and meningococci.

One of the earliest recognized virulence factors is the capsule, which allows microorganisms to resist phagocytosis. The complement component C3b and immunoglobulin G are opsonins (that is,

serum factors promoting the attachment and ingestion of microorganisms by phagocytic cells) when deposited on the microbial cell surface. The C3b and antibody are recognized by receptors in the plasma membrane of the phagocyte. In encapsulated bacteria, however, immunoglobulin G and C3b are deposited on the cell wall of the organism and are thus covered over by the capsule. This precludes their interaction with receptors on the phagocytes, and the microorganism avoids recognition and ingestion. Such a mechanism operates in some strains of *Staphylococcus aureus*, *Streptococcus pneumoniae*, and the yeast *Cryptococcus neoformans*, and may apply to other encapsulated bacteria. Antibodies directed against capsular antigens are required for efficient phagocytosis of encapsulated bacteria.

Certain bacteria, although efficiently phagocytized, are able to resist the intracellular killing action of phagocytes. In some of these cases, resistance to killing has been associated with cell-surface components (for example, a peptidoglycolipid in *Mycobacterium* species).

Specific host defenses. Some microorganisms can alter the surface that they present to the host. This antigenic variation is best understood in trypanosomes, which are protozoa. The trypanosome surface coat is composed of a protein known as the variant-surface glycoprotein to which the host develops immunity. The host defense system fails to eliminate the parasites completely because a few trypanosomes replace their coat with a different glycoprotein, requiring a new immunological response, by a process known as antigenic variation. The disease thus shows a cyclical nature, and antigenically different forms are isolated from patients in relapse. *See* CELLULAR IMMUNOLOGY; PHAGOCYTOSIS; RELAPSING FEVER; TRYPANOSOMATIDAE.

Host damage. Bacterial substances causing host damage can be divided into endotoxins, exotoxins, and tissue-degrading enzymes. Lipopolysaccharide (endotoxin) has a variety of pharmacologic and pathological effects, most of which seem to reside in the part of the molecule known as lipid A. Damaging effects of lipopolysaccharide include coagulative responses, necrotic lesions, effects on circulating leukocytes and platelets, changes in blood pressure, and fever-producing activity. Peptidoglycan shares many of the pathogenic properties of lipopolysaccharide, but is generally less potent.

Both peptidoglycan and lipopolysaccharide activate the complement system. Too much activation may damage the host. A variety of microbial cell-surface antigens can interact with antibodies and complement to form circulating immune complexes that may damage host tissues.

Exotoxins are proteins secreted by growing bacteria. Compared to endotoxins, they are less stable to heat, more specific in their actions, and more potent. Exotoxins may be divided into two major groups, membrane-damaging (phospholipases, hemolysins, lysins) or intracellular-acting. Phospholipases are produced by species of *Clostridium* and hemolysins by *Staphylococcus aureus* strains. *Staphylococcus aureus* also produces leukocidin, which creates pores in the cell membranes of polymorphonuclear leukocytes and macrophages. Examples of intracellular-acting toxins are *Corynebacterium diphtheriae* toxin, which interferes with protein synthesis and is lethal for many animal species; *Vibrio cholera* toxin, an enterotoxin causing massive diarrhea and dehydration; and *Clostridium tetanus* and *C. botulinum* toxins, which are potent neurotoxins. *See* BOTULISM; CHOLERA; DIPHTHERIA; TETANUS.

Certain generalizations can be made about these toxins. The genes specifying their production are often encoded by a prophage (a virus integrated into the bacterial chromosome) or by a plasmid. These toxins typically show a dichain construction: the alpha chain possesses the biological activity, while the beta chain mediates binding to receptors on susceptible host cells. These host receptors may actually be hormone receptors also used by the toxins. Toxins are often produced in inactive proenzyme forms that must be cleaved enzymatically to release active alpha fragments. These fragments may inhibit protein synthesis (such as in diphtheria toxin and *Pseudomonas aeruginosa* exotoxin A) or may elevate cellular cyclic adenosine monophosphate levels, altering normal ion flux (such as in cholera toxin).

A variety of bacteria secrete enzymes such as lipases, proteases, coagulases, hyaluronidases, and nucleases. These may play a role in virulence by degrading host tissues and perhaps helping nourish the parasite. In most cases, it has not been possible to make definitive statements about the role of these substances as virulence factors. *See* TOXIN.

From an evolutionary standpoint, the host-parasite relationship can be regarded as an uneasy truce. If the parasite acquires excessive virulence and kills its host, it seriously diminishes its chances of finding a new host. Furthermore, much of the apparent virulence of pathogens is actually a reflection of the reaction of the host's defense systems. *See* MEDICAL BACTERIOLOGY.

Brian Wilkinson

Bibliography. G. H. Cassell (ed.), Microbial surfaces: Determinants of virulence and host responsiveness, *Rev. Infect. Dis.* 10(suppl. 2):S273–S456, 1988; B. B. Finlay and S. Falkow, Common themes in microbial pathogenicity, *Microbiol. Rev.*, 53:210–230, 1989; M. Schaechter, G. Medoff, and D. Schlessinger, *Mechanisms of Microbial Disease*, 3d ed., 1998.

Virus

Any of the elementary agents that possess some of the properties of living systems, such as having a genome and being able to adapt to changing environments. However, viruses are not functionally active outside their host cells. Viruses share three characteristics: (1) their simple, acellular organization consisting of a nucleic acid genome surrounded by a protective protein shell, which may itself be enclosed within an envelope that includes a membrane; (2) the presence of either DNA or RNA, but

not both; and (3) their inability to reproduce independent of host cells. In essence, viruses are nucleic acid molecules, that is, genomes that can enter cells, replicate in them, and encode proteins capable of forming protective shells around them.

One debate since the discovery of the chemical nature of viruses is whether or not these agents are living microorganisms. Lacking any metabolism, they can be considered as lifeless as isolated chromosomes. They are, however, capable of reproducing their own kind abundantly and, in that sense, possess one of the major attributes of life. Some scientists believe that this is sufficient reason to term them as living. Others, however, believe that these agents should not be referred to as living, as they lack many of the basic abilities of life. It is preferable to refer to them as functionally active or inactive rather than living or dead.

The origin and evolution of viruses is a subject of much speculation among virologists. Two hypotheses have been proposed to account for their development. The first is that some of the more complex viruses, such as the poxviruses and herpesviruses, arose from small cells (probably prokaryotic) that parasitized larger, more complex cells. These parasitic cells then became simpler and more dependent on their hosts in a process known as retrograde evolution. There are several problems with this hypothesis. For example, viruses are radically different from prokaryotes; intermediate forms between the prokaryote and the complex viruses should still be in evidence but have not been found. The second hypothesis is that viruses are cellular nucleic acids that have become partially independent of the cell. Evidence that suggests this hypothesis may be true includes the observation that the nucleic acids of retroviruses contain sequences quite similar to those of normal cells.

Viruses are recognized as significant causes of disease in animals and plants. Many of the most important diseases that afflict humankind, including poliomyelitis, hepatitis, influenza, the common cold, measles, mumps, chickenpox, herpes, rubella, hemorrhagic fevers, encephalitis, and the acquired immunodeficiency syndrome (AIDS), are caused by viruses. AIDS, first diagnosed in 1981 and caused by the human immunodeficiency virus (HIV), represents an unprecedented epidemic with enormous implications for world health because case mortality rates for HIV-infected individuals exceed 90%. Current estimates of the numbers of infected human beings (intravenous drug abusers, male and female sex partners of HIV-positive individuals, and children of infected women) approach 45 million, and their numbers are still increasing rapidly in many parts of the world. So far, efforts to prevent or control HIV infection have been largely unsuccessful. Viruses also cause diseases in livestock (foot-and-mouth disease, avian influenza, hog cholera, distemper) and plants (tobacco mosaic disease, tomato ring spot, rice dwarf disease) that are of great economic importance. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); PLANT VIRUSES AND VIROIDS.

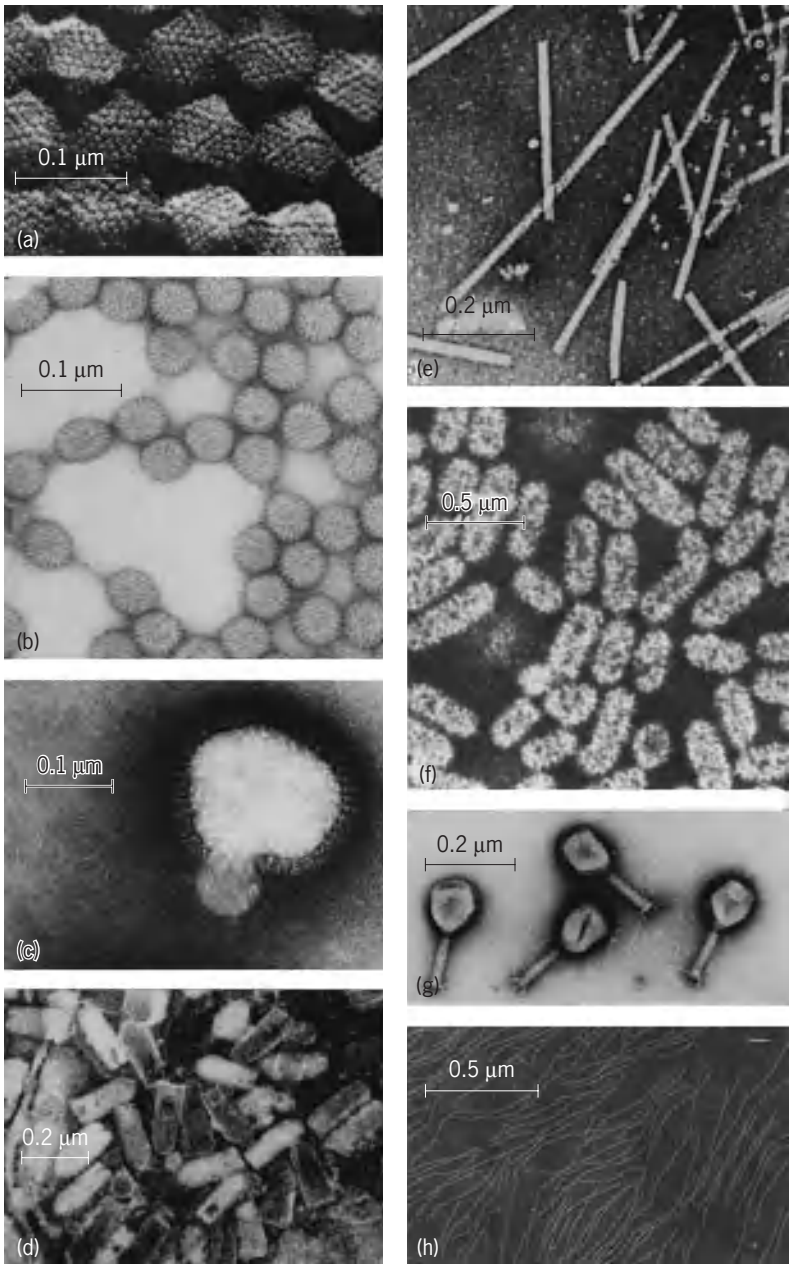
Viruses are also the simplest model systems for studying basic problems in biology. Their genomes are often no more than one-millionth the size of, for example, the human genome; yet the principles that govern the behavior of viral genes are the same as those that control the behavior of human genes. Viruses thus afford unrivaled opportunities for studying mechanisms that control the replication and expression of genetic material. *See* HUMAN GENOME PROJECT.

Morphology and size. Virus particles range in size from about 10 to 400 nanometers in diameter. The smallest viruses are about the size of a ribosome, while the largest, most complex viruses (such as vaccinia) are visible in the light microscope. However, most viruses must be visualized using the electron microscope. Although viruses differ widely in shape and size (see **illustration**), they are constructed according to certain common principles. Basically, viruses consist of nucleic acid and protein arranged in a structure known as the nucleocapsid. The nucleic acid is the genome containing the information necessary for virus multiplication and survival; the protein is arranged around the genome in the form of a layer or shell that is termed the capsid. Some viruses consist only of a naked nucleocapsid, while in others the nucleocapsid is surrounded by a membrane on the outside of which "spikes" composed of glycoproteins may be attached; this is termed the envelope. The complete virus particle is known as the virion, a term that denotes both intactness of structure and the property of infectiousness.

There are four general morphological types of capsids and virion structure. Icosahedral capsids are regular polyhedrons with 20 equilateral triangular faces and 12 vertices, appearing as spherical shapes when viewed at low power on the electron microscope. Helical capsids are shaped like hollow cylinders containing an extended viral nucleic acid. Enveloped viruses contain either helical or icosahedral capsids and have a roughly spherical shape overall. Complex viruses are neither purely icosahedral or helical and possess tails or other structures not found in simpler viruses.

Viral genomes. Viral genomes are astonishingly diverse. Some consist of deoxyribonucleic acid (DNA), others of ribonucleic acid (RNA). Some are double-stranded, others single-stranded. Some are linear, others circular. Some contain positive-sense RNA, meaning the genome can be directly read and translated into proteins, while others consist of negative-sense RNA and must be converted to a positive strand in order to be translated. Some consist of one molecule, others of several (up to 12). Their size also varies within wide limits: they range from 3000 to 280,000 base pairs if double-stranded, and from 5000 to 27,000 nucleotides if single-stranded. *See* DEOXYRIBONUCLEIC ACID (DNA); RIBONUCLEIC ACID (RNA); VIRUS CLASSIFICATION.

Viral proteins. Viral genomes encode three types of genetic information. First, they encode the structural proteins of virus particles. Second, most viral genomes can be neither replicated nor decoded (that



Electron micrographs of highly purified preparations of some viruses. (*a*) Adenovirus. (*b*) Rotavirus. (*c*) Influenza virus (*courtesy of George Laser*). (*d*) Vesicular stomatitis virus. (*e*) Tobacco mosaic virus. (*f*) Alfalfa mosaic virus. (*g*) T4 bacteriophage. (*h*) M13 bacteriophage.

is, transcribed into messenger RNA) by host-cell enzymes. Most viruses, therefore, encode enzymes capable of transcribing their genomes into messenger RNA molecules that are then translated by host-cell ribosomes, as well as nucleic acid polymerases capable of replicating their genomes. For example, the RNA genome of retroviruses encodes the enzyme reverse transcriptase, which is carried within the capsid when the virus infects a host cell. This enzyme allows these viruses to copy their genome and insert it into the host-cell chromosome, where it remains for long periods of time before it becomes activated by an outside stimulus. Many viruses also encode nonstructural proteins with catalytic and other functions necessary for virus particle maturation and morphogenesis. Third, many viruses encode proteins that interact with components of host-cell defense mechanisms against invading infectious agents. The more successful these proteins are in neutralizing these defenses, the more virulent viruses are and the more severe the resulting disease. Larger viruses encode specific proteins with these functions; smaller viruses do not, and for them these functions are exercised primarily by the structural proteins that they encode.

Virus–cell interaction. The two most commonly observed virus–cell interactions are the lytic interaction, which results in virus multiplication and lysis of the host cell, and the transforming interaction, which results in the integration of the viral genome

into the host genome and the permanent transformation or alteration of the host cell with respect to morphology, growth habit, and the manner in which it interacts with other cells.

Lytic interaction. The lytic virus–cell interaction is best thought of in terms of a cycle, the one-step growth cycle. This cycle involves the sequential, precisely regulated expression of the information encoded in the viral genome. The manner in which viral genomes express the information encoded in them is characteristic of each virus family.

The key features of the one-step growth cycle, which lasts, depending on the virus, from 6 to 36 hours, are as follows. The parental virus particle adsorbs (attaches) to specific receptors located on the host-cell surface, and is internalized by a process akin to phagocytosis. The viral genome is then either completely or partially released and expresses the information that it encodes by being transcribed into messenger RNA molecules (this is not necessary if the genome is positive-sense RNA), which are then translated. At the same time, the viral genome replicates either in the nucleus or in the cytoplasm, depending on the virus. Generally, DNA viruses replicate their genome in the nucleus, whereas RNA viruses carry out this process in the cytoplasm. When a sufficient amount of capsid proteins has accumulated, morphogenesis proceeds and progeny virus particles (up to 10^6 per cell for small viruses) are formed. Throughout this period, degradative and necrotic changes are elicited that result in the disintegration (lysis) of the cell. Naked viruses are released when the cells lyse; enveloped viruses are released when their nucleocapsids bud through the outer cell membrane, thereby acquiring their envelope.

When viruses infect multicellular organisms, the one-step growth cycle is repeated many times until, for one reason or another, further multiplication is arrested, or the host dies. The symptoms of virus infection vary widely, from asymptomatic infections detectable only by the formation of antibodies, to progressively more severe disease culminating in death. For every virus there are variant strains that differ in the severity of their effects on cells and host organisms: the more severe the effects, the more virulent the virus strain is said to be.

Transforming interaction. Certain viruses interact with cells not only by means of the lytic interaction but also by means of an interaction in which virus multiplication is repressed and the host cell is not destroyed. In this type of interaction, either the viral genome is integrated into the genome of the host cell or it replicates as a plasmid. The cells transformed by these viruses do not die and are capable of multiplying. Transformed bacterial cells are said to be lysogenic; they grow like normal cells, but when the integrated viral genome that they harbor is activated, which happens with a frequency of about 1 in 10^5, a viral growth cycle ensues and the cells lyse. Transformed animal and plant cells are also capable of multiplying; they often grow into tumors, and the viruses that cause such transformation are known as tumor viruses.

Tumor viruses are extremely important in studies on the mechanisms of cancer induction (oncogenesis) for two reasons. First, infection with tumor viruses generally transforms normal cells into potential tumor cells very efficiently; with all other tumorigenic agents, the frequency of transformation is much lower. As a result, studies at the biochemical and molecular level are possible on virus-induced transformation but not on transformation caused by other agents. Second, elucidation of the mechanisms by which tumor viruses transform cells has identified the causes of carcinogenesis in molecular terms.

Viruses can cause cancer in several ways. A virus may bring oncogenes into a cell and insert them into its genome. For example, the retrovirus Rous sarcoma virus carries a gene that codes for tyrosine kinase. This enzyme is responsible for the phosphorylation of many cellular proteins which are regulated by the addition of a phosphate group. By bringing in a new source of phosphorylation, the regulation of these proteins is altered and cell growth and behavior change as a result, potentially leading to cancer.

Other oncogenic viruses carry one or more very effective promoters, which can then insert next to a cellular oncogene. This promoter then stimulates transcription of the oncogene, leading to cancer. For example, some chicken retroviruses cause lymphomas when their genome is integrated next to a cellular oncogene. This mechanism is known as insertional activation or proviral insertion.

A final mechanism involves the inactivation of proteins that negatively regulate cell division. The normal regulation of cell division involves the extremely sensitive interplay of proteins that either promote or inhibit it, depending on the signals received. Certain viruses, primarily DNA-containing viruses, encode proteins that bind or inactivate proteins that inhibit cell division. These viral proteins can therefore result in the activation of cell division, promoting formation of cancer cells as well as stimulating virus reproduction. *See* CANCER (MEDICINE); ONCOLOGY; RETROVIRUS; TUMOR VIRUSES.

Antiviral chemotherapy. Because viruses enter host cells and make use of host-cell enzymes and constituents to reproduce, development of drugs to treat viral infections seemed a remote possibility for many years. A drug that would block virus reproduction would likely be toxic to the host. However, inhibitors of virus-specific enzymes and life-cycle stages have now been discovered. Most antiviral drugs in use today disrupt either viral nucleic acid synthesis or specific stages in the virus life cycle. Viral nucleic acid synthesis is almost always carried out by virus-encoded enzymes that do not exist in uninfected cells and are therefore excellent targets for antiviral chemotherapy. Stages in the virus life cycle that have been elucidated through research are also used as targets in a number of antiviral treatments; these include enzymes required for attachment to host cells and enzymes responsible for virus maturation and release.

Numerous chemical compounds have been described that inhibit the multiplication of viruses.

Only a few, however, inhibit virus multiplication efficiently in the body without undesirable side effects. The most successful of these compounds are analogs of ribo- and deoxyribonucleosides, such as iododeoxyuridine, trifluorothymidine, and adenosine arabinoside (vidarabine, ara-A) for the topical treatment of herpes simplex keratitis and encephalitis; acyclovir and gancyclovir for the treatment of herpesvirus and cytomegalovirus infections, respectively; zidovudine (AZT), lamivudine (3TC), didanosine (ddI), zalcitabine (ddC), and stavudine (d4T) for the treatment of HIV infections (AIDS); and ribavirin (Virazole®) for the treatment of respiratory syncytial virus and Lassa fever virus infections. Another viral function that has been targeted is the cleavage of polyproteins, precursors of structural proteins, to their functional components by virus-encoded proteases; examples include the HIV protease inhibitors (saquinvir, indinavir, and ritonavir) used for treatment in AIDS patients. *See* CHEMOTHERAPY; CYTOMEGALOVIRUS INFECTION; HERPES; INFLUENZA; RESPIRATORY SYNCYTIAL VIRUS; VIRUS CHEMOPROPHYLAXIS.

Other promising new approaches to antiviral chemotherapy include the use of agents that interfere with virus adsorption (primarily against HIV); agents that inhibit the enzyme neuraminidase, which is essential for the entry of influenza virus into cells and the subsequent release of its progeny from cells; agents that bind in hydrophobic pockets of viral capsids, thereby stabilizing them and interfering with uncoating (primarily against picornaviruses such as rhinoviruses); the targeted introduction of toxins into infected cells (against viruses for which the receptors are known, such as HIV, rhinoviruses, and Epstein-Barr virus); and the introduction into cells of specific antisense RNA sequences.

Antiviral agents on which much interest is focused are the interferons. Interferons are cytokines or lymphokines that regulate cellular genes concerned with cell division and the functioning of the immune system. Their formation is strongly induced by virus infection; they provide the first line of defense against viral infections until antibodies begin to form. Interferons interfere with the multiplication of viruses by preventing the translation of early viral messenger RNAs. As a result, viral capsid proteins cannot be formed and no viral progeny results. Interferon is clinically useful in the treatment of hepatitis B and C infection.

By far the most effective means of preventing viral diseases is through mobilization of the immune system by vaccines. There are two types of antiviral vaccines: inactivated and attenuated active. Most of the antiviral vaccines currently in use are of the latter kind. The principle of antiviral vaccines is that inactivated virulent or active attenuated virus particles cause the formation of antibodies that neutralize a virulent virus when it invades the body. *See* ANIMAL VIRUS; VACCINATION; VIRUS, DEFECTIVE.

W. K. Joklik; Marcia M. Pierce

Bibliography. L. Collier and J. Oxford, *Human Virology*, 2d ed., Oxford University Press, 2000; D. M. Knight and P. M. Howley (eds.), *Fundamental Virology*, 4th ed., Lippincott, Williams & Wilkins, Philadelphia, 2001; L. M. Prescott, J. P. Harley, and D. A. Klein, *Microbiology*, 6th ed., McGraw-Hill, New York, 2005; V. van Regenmortel et al. (eds.), International Committee on Taxonomy of Viruses, *Virus Taxonomy: Classification and Nomenclature of Viruses: Seventh Report of the International Committee on Taxonomy of Viruses*, 2000; S. Schlesinger and M. Schlesinger, Viruses, in J. Lederberg (editor-in-chief), *Encyclopedia of Microbiology*, 2d ed., vol. 4, pp. 796–810, Academic Press, San Diego, 2000; B. A. A. Voyles, *The Biology of Viruses*, 2d ed., McGraw-Hill, Chicago, 2002.

Virus, defective

A virus that by mutation has lost the ability to be replicated in the host cell without the aid of a helper virus. The virus particles (virions) contain all the viral structural components; they can attach, penetrate, and release their nucleic acid ribonucleic acid (RNA), deoxyribonucleic acid (DNA)] within the host cell. However, since the mutation has destroyed an essential function, new virions will not be made unless the cell was simultaneously infected with the helper virus, which can provide the missing function. Only then will the cell produce a mixed population of new helper and defective viruses. Occasionally, when their nucleic acids become integrated in the DNA of the host cell, defective viruses persist in nature by propagation from mother cell to daughter cell.

Deletion mutants. The most important group of defective viruses are deletion mutants. They are derived from their homologous nondefective (wild-type) virus through errors in the nucleic acid replication, which result in the deletion of a fragment in the newly synthesized molecules. The deletion is a consequence of an excision or of a failure in the synthesis of the fragment. The defective nucleic acid must be capable of self-replication, at least in the presence of the wild-type virus, and must combine with other viral components to form a particle in order to exit the cell. Any other types of deleted nucleic acids, although probably transiently formed, are dead-end products, undetectable among the viral progeny.

The defective RNA tumor viruses are deletion mutants. Mammalian and most avian sarcoma viruses require a nondefective leukemia virus as a helper virus. Usually the specificity for a certain type of host cell exhibited by the defective virion depends on the helper virus, indicating that one of the virion surface proteins has been furnished by the helper virus gene. These proteins are involved in interactions with cellular surface receptors, and thus determine whether a cell can serve as a host for viral infection. There are also defective leukemia viruses sometimes called acute leukemia viruses. In all cases of defective RNA tumor viruses, the deleted RNA fragment was replaced by a new segment, which carries genetic

information required for the induction of tumors or neoplasia in the host. This gene, the oncogene, is thought to have arisen by a modification of a normal cellular gene. *See* ONCOLOGY; TUMOR VIRUSES.

Defective interfering particles. Certain deletion mutants, which interfere with the replication of the helper virus, are called defective interfering particles. Cells infected with a mixture of a standard virus and its defective interfering particles produce much less of the standard virus than the same type of cells infected with the standard virus in the absence of defective interfering particles. Defective interfering particles were discovered by P. Von Magnus in 1947 in infections with influenza virus. He called them incomplete or immature particles, and their ability to attenuate the replication of influenza virus he termed autointerference. Since then, defective interfering particles have been found in most of the major groups of viruses (see **table**).

The widespread occurrence of defective interfering particles has prompted the suggestion that they must perform an important role in the evolution of viral infection. The self-limiting nature of the infection imposed by the generation of defective interfering particles must be advantageous to the virus. Since viruses require living host cells for self-replication, an unlimited spread of infection resulting in the death of all available host cells is clearly not expedient to the survival of the virus. The plausibility of this hypothesis is borne out by the observation that in Australia, during attempts to control the rabbit population with myxomavirus, a rapid drift from the original virulent virus strain to less virulent strains has taken place.

Occurrence of defective interfering particles in RNA viruses*

Virus group	Member name
Negative strand	
Rhabdo	Vesicular stomatitis
	Rabies, bovine ephemeral fever, others
Paramyxo	Sendai
	Newcastle disease
	Measles
	Mumps
Orthomyxo	Influenza
	Fowl plague
Arena	Lymphocytic choriomeningitis, Amapair, Parana, Pichinde
	Tacaribe
Bunya	Bunyamwera
	LaCrosse
	Turlock
Positive strand	
Picorna	Poliovirus
	Mengovirus
Calici	Feline calcivirus
Toga	Sindbis, Semliki Forest, West Nile, Japanese encephalitis
Corona	Mouse hepatitis
Double strand	
Reo	Reovirus
Orbi	Bluetongue virus of sheep
Fungal	Yeast-killer, others
Other	Infectious pancreatic necrosis virus

*Defective interfering particles were also observed in DNA viruses of the Papova and Herpes groups. Only one type of adenovirus (adeno-12) seems to generate these particles.

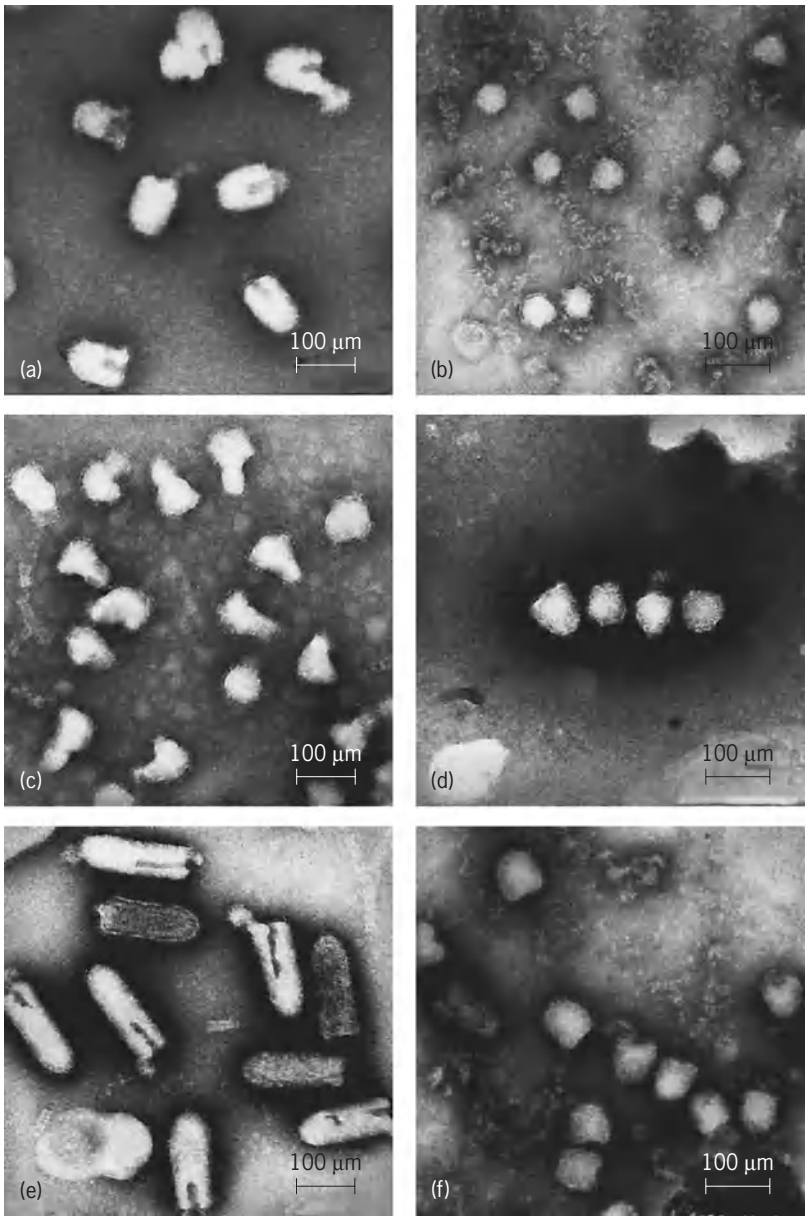
SOURCE: J. Perrault, *Curr. Topics Microbiol. Immunol.*, vol. 93, 1981.

Morphology. Defective interfering particles derived from "spherical" (icosahedral) viruses, when viewed with the electron microscope, are indistinguishable from the standard virion. Since the nucleic acid of these virions is buried in the interior of the spherical structure, the deleted nucleic acid of the defective interfering particle has very little effect on the overall particle dimensions. On the other hand, in elongated (rod or bullet-shaped) virions, structural proteins are wound around the backbone of the ribbonlike nucleic acid molecule, whose size determines the long dimension of the particle. Consequently, in these cases defective interfering particles are shorter than their homologous virions, the foreshortening of the particle being proportionate to the size of the deletion. This is illustrated in the electron micrographs of vesicular stomatitis virus and some of its particles shown in the **illustration**. A nucleic acid with a 50% deletion is contained in a defective interfering particle (illus. *a*), half the length of the homologous virions. Other, more extensive deletions are found in particles correspondingly shorter.

Genome structure. A defective interfering particle's requirements for genome replication and incorporation into a virionlike particle result in the selection of certain nucleic acid structures. In the case of DNA viruses, the selection leads to deleted DNAs, with tandem duplications of a segment which contains the origin of replication. One advantage of these structures is their more efficient replication through multiple initiations, and another may reside in a selection of a minimum DNA size, in order to form a particle. The minimum DNA size is sometimes achieved by acquisition of host-cell DNA segments. The mechanism by which these structures are formed is not understood, although many hypotheses have been proposed.

Based on the nucleic acid structure, RNA-containing defective interfering particles can be broadly classified into two groups. In one group are particles in which the deletion of the linear RNA molecule took place at one end. In the other group the deletion was internal with conserved ends. Some viruses, such as vesicular stomatitis virus, generate both types of defective interfering particles. The terminally deleted RNAs have yet another structural peculiarity: a sequence of approximately 50 terminal nucleotides (the building blocks of nucleic acids) at the two ends are inverted copies of each other (inverted terminal repeats). Such structures are not found in the homologous viral genome, and are thought to be intimately connected with the ability of these defective interfering particles to interfere with the replication of the virus. They also furnish important clues in studies of the molecular mechanisms by which these particles are generated.

Interference. It is generally thought that the reason defective interfering particles interfere with the production of their homologous viruses is that the



Electron micrographs of virion and defective interfering particles of vesicular stomatitis virus. (*a*) Defective interfering particle with a 50% deleted RNA. (*b*) Defective interfering particle with an 85% deleted RNA. (*c*, *d*, *f*) Defective interfering particles from various isolates, all with a 70% deleted RNA. (*e*) Vesicular stomatitis virion. (*From M. E. Reichmann, C. R. Pringle, and E. A. C. Follett, J. Virol., 8:156, 1971*)

nucleic acids of defective interfering particles have a replicative advantage over those of the homologous viruses and are capable of competing successfully for the synthetic enzymes. If at some stage of the infectious cycle these enzymes are in limited supply, they might be expected to be engaged predominantly in the synthesis of the defective-interfering-particle nucleic acid. In that case, insufficient enzyme will be available for viral nucleic acid replication. In the synthesis of defective-interfering-particle DNA with reiterated origins of replication, multiple initiation processes might tie up most of the available enzymes. Alternatively, in the case of terminally deleted defective-interfering-particle RNA, the unique inverted repeats may represent a stronger binding site for the synthetic enzymes than the corresponding but different structures of the virion RNA.

Defective interfering particles with RNAs deleted internally seem to interfere by a different mechanism. In repeated rounds of replication, a selection of defective interfering particles with progressively shorter RNA molecules has been reported. This observation suggested that the rate of RNA synthesis might be the major factor in interference. The shorter defective-interfering-particle RNA molecules, which take less time to be synthesized than the longer virion RNA, would accumulate if the rate of synthesis were the major time-determining factor. In multiple rounds of infections, the shortest defective-interfering-particle RNAs would eventually outnumber both virion RNAs and the longer defective-interfering-particle RNAs, and utilize all the available synthetic enzymes. *See* VIRUS INTERFERENCE.

Specificity of helper and interference functions. Defective interfering particles can interfere only with the cellular synthesis of the virus strain from which they were originally derived, or with virus strains closely related to it. The more distant the relationship, the weaker the interference. The extent of interference can therefore serve as a measure of virus strain relatedness, in addition to other classification criteria such as morphology and serology. Similarly, the ability to serve as a helper to a defective virus is highly specific. In some cases of defective viruses (unlike the defective-interfering-particle group), the relationship between the helper virus and defective virus is not obvious. The defective adeno-associated virus and the satellite tobacco necrosis virus are examples in which the helper-defective virus relationship is also specific. Only adenoviruses and tobacco necrosis viruses can serve as helpers, respectively.

Persistent infections. Defective viruses have been used in studies of persistent, or low-level, infections. Normally cytolytic viruses kill their host cells both in cultures and in intact organisms. Occasionally, however, some cells survive, multiply, and simultaneously produce low levels of virus. In whole organisms, persistent infections can result in periodic episodes of the acute viral disease, especially if the immune response to the virus becomes temporarily compromised. Persistent infections are also involved in degenerative brain diseases and can lead to complications, sometimes years after a bout with an acute virus disease, such as measles. Laboratory models of carrier cultures infected with several viruses have been developed. In some of these the aid of defective interfering particles in the attenuation of the virulence was found to be required for both the establishment and maintenance of the culture. Although defective interfering particles were isolated also in a few cases from naturally occurring persistent infections, it is not clear whether they play a major role in the evolution of these diseases. In many cases the diminished virulence observed in persistent infections is the property of the infectious virion itself, rather than being due to interference by defective interference particles. Frequently virions isolated from persistent infections are temperature-sensitive mutants.

Temperature-sensitive mutants. A mutation might lead to a change of a single amino acid of a protein

encoded in the viral genome. Such mutations usually precipitate a change in the three-dimensional structure of the pertinent protein, accompanied by a loss of biological activity. In some cases, however, the altered protein can maintain its original conformation at slightly lower than the normal body temperatures (for example, 31°C or 87.8°F instead of 37°C or 98.6°F). Virus mutants of this kind, called temperature-sensitive (ts) mutants, do not grow efficiently at normal body temperatures, but they grow like wild-type virus at the lower temperatures. These mutants belong to the group of conditional lethal mutants, which have played an important role in genetic studies. For research purposes they can be grown to good yields under the permissive conditions (such as low temperatures in the case of temperature-sensitive mutants), while other types of mutants often cannot be obtained in adequate quantities.

The reduced growth and virulence of temperature-sensitive mutants at normal body temperatures might serve as a basis for their selection in persistent infections. In any case, temperature-sensitive mutants have been isolated from such infections in animals or carrier cultures of foot-and-mouth disease, sendai, vesicular stomatitis, herpes, influenza, coxsackie, and some alpha viruses. It is not known whether defective interfering particles, temperature-sensitive mutants, or both constitute a major path to persistent infections and virus latency. *See* ANIMAL VIRUS; MUTATION; VIRULENCE; VIRUS INFECTION, LATENT, PERSISTENT, SLOW; VIRUS INTERFERENCE. M. E. Reichmann

Bibliography. H. Fraenkel-Conrat and R. R. Wagner (eds.), *Comprehensive Virology*, vol. 10, 1977, vol. 16, 1980, vol. 18, 1983; J. A. Levy, H. Fraenkel-Conrat, and R. A. Owens, *Virology*, 3d ed., 1995; M. E. Reichmann and W. M. Schnitzlein, Defective interfering particles of rhabdoviruses, *Curr. Topics Microbiol. Immunol.*, 86:123–168, 1979.

Virus chemoprophylaxis

The treatment of viral disease by chemical drugs has been approached experimentally through the inhibition of intracellular metabolic processes that otherwise lead to the synthesis of viral constituents. Compounds have also been tested for their capacity to inhibit adsorption, penetration, or release of infectious virus from the host cell. In order to be considered for chemoprophylaxis, a compound must have a greater specificity for inhibiting virus-directed reactions than for normal host cell reactions. Although some success has been achieved, there are no inhibitors available that can be used in the routine treatment of viral infections in humans.

Very few inhibitors have been found which have a specific effect on virus penetration or release from the host cell. Amantadine, a symmetrical amine, inhibits specifically certain members of the myxovirus group through blocking viral penetration into the host cell. When administered prophylactically, that is, in advance of exposure to the virus, amantadine is reported to have a significant protective effect in experimental animals and humans against certain influenza A strains but not against influenza B or other viruses. This protection consists mainly of a modification of the disease to a milder form.

Guanidine and 2-(α-hydroxybenzyl)-benzimidazole (HBB) specifically inhibit the replication of many picornaviruses in the test tube, for example, poliovirus, by restricting the formation of virus-specific enzymes, which indirectly results in an inhibition of the synthesis of viral ribonucleic acid (RNA) and coat proteins. In experimentally infected animals, there is no protective effect by either inhibitor. This is probably due to rapid production of drug-resistant mutants.

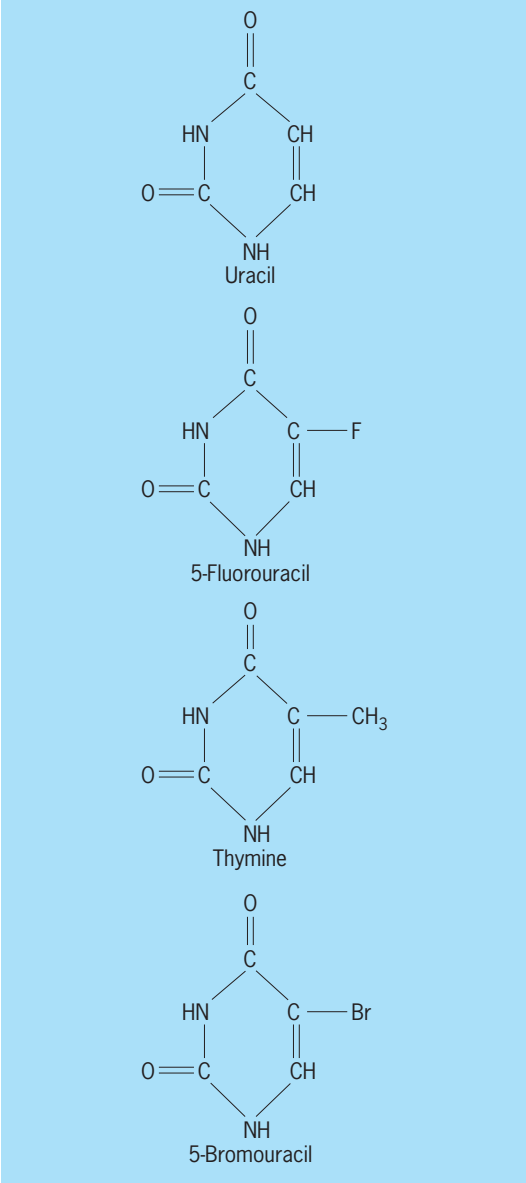
Methisazone (isatin-β-thiosemicarbazone) is an inhibitor of many members of the poxvirus and adenovirus groups. Treatment of infected human cells with this inhibitor results in the formation of immature, noninfectious virus particles. A derivative of this compound has been shown to be an effective prophylactic for smallpox in humans if given within 24–48 h after exposure.

Many purine and pyrimidine analogs inhibit both ribonucleic acid (RNA) and deoxyribonucleic acid (DNA) synthesis. By the addition of ribose or deoxyribose to make the corresponding riboside or deoxyriboside, the activity of the analog can be directed preferentially toward the inhibition of RNA or DNA. The significance of this structural specificity of the sugar component for the inhibition of virus synthesis is evident in the finding that the riboside of 5-fluorouracil is a much more effective inhibitor of the growth of tobacco mosaic virus than its deoxyriboside. Ribosides of halogenated benzimidazoles are more selective inhibitors of influenza virus or poliovirus multiplication than the free benzimidazoles or their deoxyriboside. *See* NUCLEIC ACID; PYRIMIDINE.

Another difference between the two types of nucleic acid is in their base composition, since uracil is present in RNA and thymine in DNA. Pyrimidine analogs may therefore exert an action according to their structural similarity to either of these two bases (see **illus.**). With halogenated pyrimidines, the size of the halogen substituent determines the character of the compound. The van der Waals radii of fluorine and bromine resemble closely those of hydrogen and the methyl group, respectively: The size and shape of 5-bromouracil is very similar to that of thymine, and that of 5-fluorouracil is similar to that of uracil.

5-Bromouracil, especially in the form of its deoxyriboside, is an effective inhibitor of the synthesis of DNA bacteriophage but is without effect on the synthesis of RNA tobacco mosaic virus. 5-Fluorouracil, on the other hand, inhibits the growth of the RNA virus, and its action can be reversed by uridine but not by thymidine.

5-Fluoro-2′-deoxyuridine (FUDR) inhibits DNA synthesis by inhibiting the enzymatic synthesis of thymidylic acid; 5-bromo-2′-deoxyuridine (BUDR) and 5-iodo-2′-deoxyuridine (IUDR) are incorporated into DNA, resulting in the production of a faulty nucleic acid which does not function normally. All three of these halogenated deoxyuridines inhibit replication of members of the major DNA virus groups: papova-, adeno-, herpes-, and poxviruses.



Pyrimidines uracil and thymine and their analogs.

Topical administration of IUDR is being used in humans in the treatment of corneal lesions due to herpes simplex virus. Systemic treatment of viral infections is not practicable because of the toxicity of the drugs. However, in cases of herpesvirus encephalitis, massive near-lethal doses (0.006 oz/lb or 400 mg/kg) of IUDR have been reported to lead to complete recovery.

Another pyrimidine analog, 1-β-*d*-arabinofuranosylcytosine hydrochloride (cytosine arabinoside), inhibits DNA synthesis, and thereby the replication of DNA viruses. As with IUDR, cytosine arabinoside inhibits cellular DNA synthesis and viral DNA synthesis about equally and therefore exhibits little viral specificity.

Dactinomycin (actinomycin D) inhibits DNA-dependent RNA synthesis and thus inhibits the multiplication of DNA viruses, but not most RNA viruses. However, dactinomycin inhibits the multiplication of some RNA-containing viruses (some myxoviruses and Rous sarcoma virus). The exact mechanism of this dactinomycin inhibition is not clear. Rifampicin (Rifampin) is an antibiotic whose action depends upon its preferential inhibition of bacterial RNA polymerase over animal-cell RNA polymerase. Since poxviruses carry their own RNA polymerase for synthesizing viral messenger RNA (needed for transcription of the viral genome), rifampicin was tried against this virus in culture and found to be very effective in inhibiting its replication. Although the exact mechanism of action of this compound is not known, it is thought to block assembly of DNA and proteins into the mature virus particle.

Although not practical as chemotherapeutic agents, protein inhibitors have been useful in the study of viral replication. For example, puromycin, cycloheximide, and ρ-fluorophenylalanine all inhibit synthesis of both viral and cell proteins. *See* BACTERIOPHAGE; CHEMOTHERAPY; VIRUS.

Joseph L. Melnick

Virus classification

There is no evidence that viruses possess a common ancestor or are in any way phylogenetically related. Nevertheless, classification along the lines of the Linnean system into families, genera, and species has been partially successful. In the early 1970s the International Committee on Taxonomy of Viruses (ICTV) was established. This article is a summary of the committee's most recent findings. Based on the organisms they infect, the first broad division is into vertebrate; algae, fungi, yeast, and protozoan; invertebrate; plant; and bacterial viruses. (However, viral families may fall into more than one of these classes.) Within these classes, other criteria for subdivision are used. Among these are general morphology: envelope or the lack of it; nature of the genome (deoxyribonucleic acid, DNA, or ribonucleic acid, RNA); structure of the genome [single-stranded (ss) or double-stranded (ds), linear or circular, fragmented or nonfragmented]; mechanisms of gene expression and virus replication (positive- or negative-strand RNA); serological relationship; host and tissue susceptibility; pathology (symptoms, type of disease). Information about many of these parameters has been available only since the 1950s, and was acquired after the introduction of tissue culture laboratory techniques.

Families and Genera of Viruses Infecting Vertebrates

The families of vertebrate viruses and the shapes and sizes of their characteristic members are presented in **Fig. 1**. Families are sometimes subdivided into subfamilies; the suffix -virinae may then be used. The subgroups of a family or subfamily are equivalent to the genera of the Linnean classification. *See* ANIMAL VIRUS.

DNA viruses. The vertebrate DNA viruses are divided into 10 families: Asfarviridae, Poxviridae, Iridoviridae, Hepadnaviridae, Herpesviridae, Polyomaviridae, Papillomaviridae, Adenoviridae, Circoviridae, and Parvoviridae.

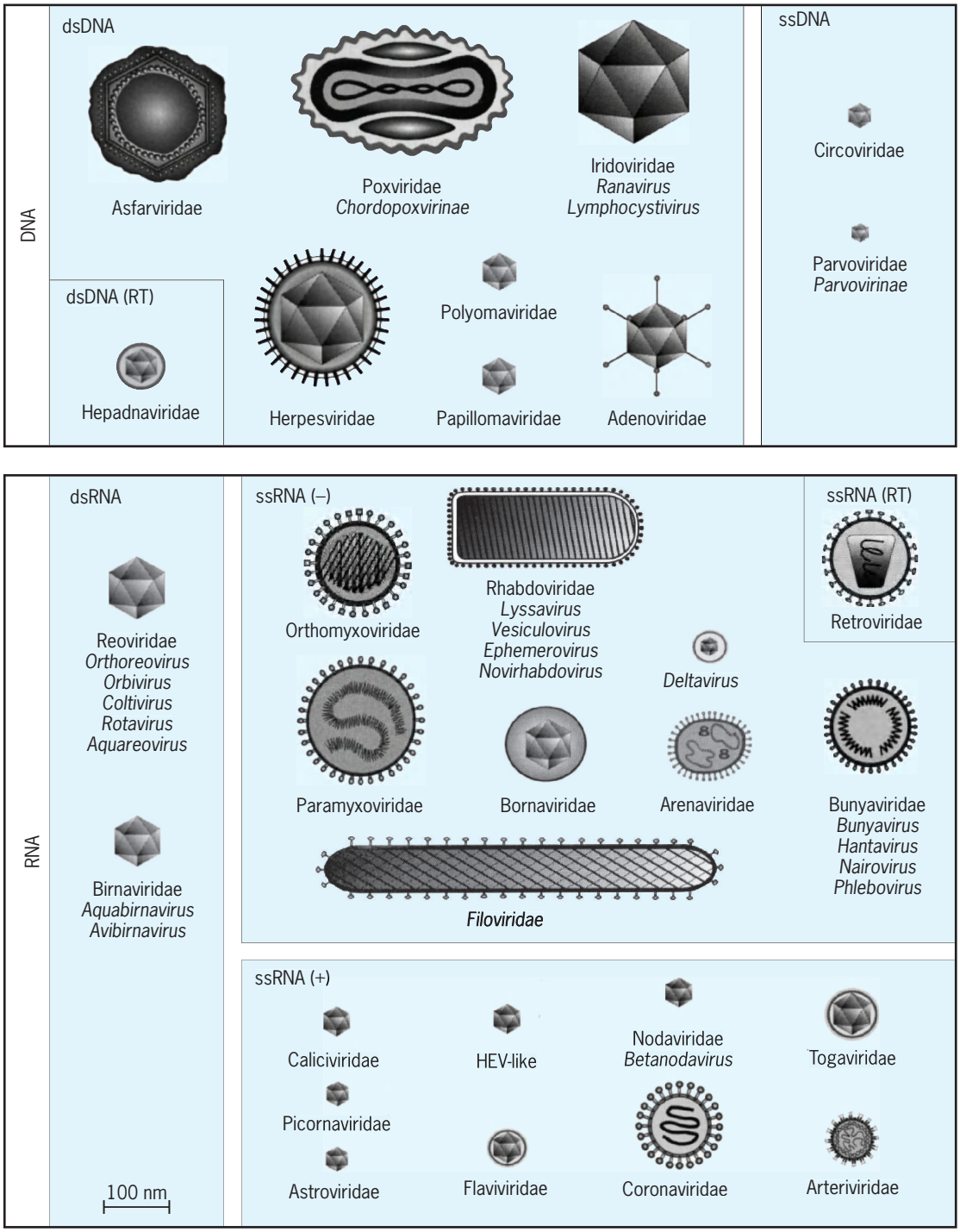


Fig. 1. Families and genera of viruses infecting vertebrates. Individual frames separate taxa of viruses containing double-stranded and single-stranded genomes. Large boxes separate taxa of viruses having DNA or RNA genomes. All diagrams have been drawn approximately to the same scale to provide an indication of the relative sizes of the viruses. (*From M. H. V. van Regenmortel et al., eds., Virus Taxonomy: 7th Report of the International Committee on Taxonomy of Viruses, p. 30, copyright 2000, with permission from Elsevier*)

Asfarviridae. The properties of this family are based on the sole genus *Asfivirus* (African swine fever virus). Virions consist of a nucleoprotein core structure (70–100 nm in diameter), surrounded by internal lipid layers and an icosahedral capsid (170–190 nm in diameter) and a lipid-containing envelope. The genome is a single molecule of linear dsDNA (170–190 kbp). African swine fever virus infects domestic and wild swine and can be transmitted by ticks, direct contact, fomites (disease-carrying objects), or mechanically by biting flies.

Poxviridae. The properties of this family are based on vaccinia virus of the *Orthopox* genus. The brick-shaped, (200 × 200 × 250)-nanometer particle contains a dumbbell-shaped nucleocapsid with a genome of dsDNA (170–250 kbp). At least 30 structural proteins have been reported. In spite of the lipid-containing envelope, most of these viruses are ether-resistant (some members of the poxviridae are, however, ether-sensitive). Generally,

members of the Poxviridae are sensitive to common detergents, formaldehyde, oxidizing agents, and temperatures greater than 40°C (104°F). The large genome encodes a great number of enzymes which lead to a more autonomous virus replication than that of any other family. Unlike any other DNA virus, the poxvirus replicates in the cell cytoplasm, independent of functions found in the cell nucleus. Poxviridae are divided into two subfamilies: Chordopoxvirinae (which infects vertebrates) and Entomopoxvirinae (which infects invertebrates; Fig. 3). Genera of Chordopoxvirinae include *Orthopoxvirus* (human smallpox), *Leporipoxvirus* (rabbit myxoma), *Avipoxvirus* (chickenpox), *Capripoxvirus* (sheeppox), *Suipoxvirus* (swinepox), *Mulluscipoxvirus* (molluscum contagiosum), *Yatapoxvirus* (Yaba monkey tumor virus), and *Parapoxvirus* (pseudocowpox). *See* SMALLPOX VIRUS.

Iridoviridae. These are mostly insect viruses except for the genus *Ranavirus* (frog viruses) and the genus *Lymphocystivirus* (fish viruses). The icosahedral virion is 120–200 nm in diameter but may be up to 350 nm, with a dsDNA genome (between 140 and 303 kbp). Several structural proteins and also lipids (not in a lipid bilayer envelope) are contained in the capsid. Viruses released from cells may have a plasma-derived outer envelope. The viruses are sensitive to treatment with ether.

Hepadnaviridae. Hepadnaviruses are spherical, sometimes pleomorphic (variably formed) viruses. The virion particles are 40–48 nm in diameter after negative staining. An outer envelope surrounds an icosahedral nucleocapsid core containing the viral genome, the viral polymerase, and associated cellular proteins. The genome consists of a single molecule of circular DNA (3.0–3.3 kbp) that is partially single stranded. The family includes two genera: *Orthohepadnavirus* (hepatitis B virus) and *Avihepadnavirus* (duck hepatitis B virus). *Hepadnavirus* infection induces overproduction of surface proteins, which are secreted as lipoprotein particles together with virus to the blood. Empty viruslike particles consisting of excess virus envelope material are present in much greater numbers than complete virions in most individuals. *See* HEPATITIS B VIRUS.

Herpesviridae. In this family the particle is 120–200 nm in diameter. Its double-stranded linear DNA (125 to 240 kbp in size) is packed in a liquid crystalline array which fills the entire volume of the icosahedral capsid (100–125 nm). This package is surrounded by a lipid bilayer envelope and several glycoproteins. The space between the icosahedral capsid and the envelope, the tegument, is filled with at least 15 distinct polypeptides. Viral infectivity is destroyed by lipid solvents and detergents. The subfamilies of this group are Alphaherpesvirinae (herpes simplex virus group), Betaherpesvirinae (cytomegalovirus group), and Gammaherpesvirinae (lymphoproliferative virus group). Epstein-Barr (EB) virus is the cause of Burkitt's lymphoma in the malaria belt and nasopharyngeal carcinoma in the Peoples Republic of China. It is also the causative agent of infectious mononucleosis. *See* CYTOMEGALOVIRUS INFECTION; EPSTEIN-BARR VIRUS; HERPES SIMPLEX VIRUS.

Polyomaviridae. This group is characterized by a nonenveloped icosahedron (72 capsomers), 40 nm in diameter, which contains five to seven unique structural proteins. Double-stranded circular DNA (5 kbp) is complexed inside the nucleocapsid to histone proteins of host cell origin. The sole genus in this family is *Polyomavirus*. Polyomaviruses include *JC polyomavirus*, which causes the rare disease known as progressive multifocal leukoencephalopathy in humans, and Simian virus 40, which causes a similar disease in rhesus monkeys.

Papillomaviridae. This family consists of nonenveloped icosahedral virions 55 nm in diameter. The capsid is composed of 72 capsomers in skewed arrangement. The genome consists of a single molecule of circular dsDNA (6800–8400 bp). The sole genus in this family is *Papillomavirus*. Species include human papillomaviruses and bovine papillomaviruses. Human papillomavirus spreads by close contact with warts. Specific human cancers, including cervical and penile cancers, have been linked to specific types of Papillomaviruses. *See* PAPILLOMAVIRUS.

Adenoviridae. This group is characterized by a nonenveloped icosahedron of 252 capsomers, 70–90 nm in diameter, which contains at least 12 unique structural proteins. The 12 vortex capsomers (pentons) exhibit one (in the mammalian group) or two (in the avian group) filamentous projections (fibers) involved in attachment to the host cell. The genome, a double-stranded linear DNA (26–45 kbp), has been sequenced for human adenovirus. The family comprises two genera, *Mastadenovirus* (mammalian, including human) and *Aviadenovirus* (avian), each containing a large number of members. *See* ADENOVIRIDAE.

Circoviridae. This family, containing the sole genus *Circovirus*, is characterized by nonenveloped icosahedral virions. The particles range in size 12–27 nm in diameter. Virions contain circular single-stranded DNA (2.0 kbp). Species in the genus include porcine circovirus and chicken anemia virus.

Parvoviridae. These virions are nonenveloped, 18–26 nm in diameter with icosahedral symmetry with 32 capsomers. Parvoviridae are the smallest viruses, containing three or four structural proteins and a ssDNA genome of 4–6 kbp. This family includes two subfamilies, Parvovirinae and Densovirinae. The subfamily includes the genera *Parvovirus*, *Erythrovirus*, and *Dependovirus*. *Dependovirus* comprises defective viruses that require a helper virus (usually an adenovirus) for their replication. The other genera, *Parvovirus* and *Erythrovirus*, are autonomous. The subfamily Densovirinae and the genus *Dependovirus* produce, without preference, particles which contain either the positive or the negative DNA strand. Parvoviruses contain only the negative strand. Densoviruses infect arthropods only and will be discussed in a subsequent section.

Single-stranded, positive-sense RNA viruses. RNA vertebrate viruses may be either single-stranded or double-stranded. The single-stranded ones are

further subdivided into positive-strand and negative-strand RNA viruses, depending on whether the RNA contains the messenger RNA (mRNA) nucleotide sequence or its complement, respectively. The RNA genes may be located on one or several RNA molecules (nonfragmented or fragmented genomes, respectively).

The single-stranded, positive-sense RNA vertebrate viruses contain eight families: Caliciviridae, Picornaviridae, Astroviridae, Flaviviridae, Nodaviridae, Coronaviridae, Togaviridae, and Arteriviridae. In addition, a genus of HEV-like viruses is found in this group but has not yet been assigned to a family.

Caliciviridae. These icosahedral particles, 27–40 nm in diameter, exhibit 32 visible cup-shaped depressions and contain 90 dimers of the major structural protein. Although their single-stranded RNA (7.4–8.3 kb) serves as mRNA, a smaller mRNA fragment has also been detected. Caliciviruses infect a broad range of animals that includes cattle, swine, and humans. Genera in this family include *Lagovirus*, *Vesivirus*, Norwalk-like viruses, and Sapporo-like viruses. The prototype of this virus group is the vesicular exanthema of swine virus (VESV).

Picornaviridae. This family contains the genera *Enterovirus* (human polio), *Cardiovirus* (mengo), *Rhinovirus* (common cold), *Hepatovirus* (hepatitis A virus), *Parechovirus* (human parechovirus), and *Aphtovirus* (foot-and-mouth disease). The nonenveloped viral capsids are icosahedral, 22–30 nm in diameter, composed of 60 identical units, consisting of three surface proteins each. Viruses contain one single-stranded RNA molecule of 7–8.5 kb in size. A small protein, VPg, is covalently bound at its 5′ end instead of the cap structure, and is polyadenylated at the 3′ end. Viruses are insensitive to ether, chloroform, or nonionic detergents. *See* ENTEROVIRUS; PICORNAVIRIDAE; RHINOVIRUS.

Astroviridae. Virions are 28–30 nm in diameter, spherical, and nonenveloped. A five- or six-pointed star is discernible on the surface of about 10% of particles. This family contains a sole genus, *Astrovirus*. A single molecule of infectious, positive-sense ssRNA (6.8–7.9 kb in size) makes up the viral genome. A poly(A) tract is located at the 3′ end of the genome. *Astroviruses* have been detected in stool samples from humans, cats, cattle, and several other animals. Astroviruses cause acute, nonbacterial gastroenteritis in children and immunocompromised adults.

HEV-like viruses. These virions are icosahedral (27–34 nm) and nonenveloped. The genome is a positive-sense ssRNA molecule of approximately 7.2 kb, with a 3′-poly(A) tail. Hepatitis E virus is associated with outbreaks and sporadic cases of fecally transmitted acute hepatitis in tropical and subtropical countries. *See* HEPATITIS E VIRUS.

Flaviviridae. These virions are 40–60 nm in diameter, spherical in shape, and contain a lipid envelope. A sole capsid protein makes up the capsid, while two or three virus-encoded membrane proteins are found in the envelope. The genome is a positive-sense ssRNA molecule, 9.6–12.3 kb in size. Genera in this group include *Flavivirus* (yellow fever virus), *Pestivirus* (bovine viral diarrhea virus), and *Hepacivirus* (hepatitis C virus). *See* YELLOW FEVER VIRUS; HEPATITIS C VIRUS.

Nodaviridae. These virions are nonenveloped, roughly spherical in shape, 29–32 nm in diameter, with icosahedral symmetry. Two ssRNA molecules (mol wt 1.1 and 0.48 $\times$ 10^6, respectively) serve as mRNAs and are both required for infectivity. Genera include *Alphanodavirus* (Nodamura virus) and *Betanodavirus* (striped jack nervous necrosis virus). All members of the Alphanodaviruses are insect viruses, although serological evidence suggests that Nodamura virus also naturally infects pigs. The Betanodaviruses were isolated from juvenile marine fish, and are associated with significant problems in commercial fish hatcheries.

Coronaviridae. Genera in this family include *Coronavirus* (severe acute respiratory syndrome or SARS) and *Torovirus* (equine torovirus). The spherical, enveloped particle is 120–160 nm in diameter, and exhibits large (20 nm) spikes projecting from the surface. The nucleocapsid is helical and is surrounded by an internal core shell of approximately 65 nm. One ssRNA molecule ranges in size from 27.6 to 31 kb. Coronaviruses are known to infect a large number of animals as well as humans. *See* SARS.

Togaviridae. This group contains the genera *Alphavirus* (Western equine encephalitis virus) and *Rubivirus* (rubella virus). The icosahedral nucleocapsid (40 nm in diameter) contains the capped and polyadenylated RNA (9–11.8 kb). The envelope contains two viral glycoprotein spikes. Many of these viruses are transmitted by insects and are the cause of human diseases, including several types of encephalitis. Rubella virus causes German measles, a benign childhood disease that causes severe fetal defects in pregnant women exposed to the virus. The viruses are spherical, enveloped particles 70 nm in diameter. *See* RUBELLA.

Arteriviridae. The sole genus in this family is *Arterivirus* (equine arteritis virus). These virions are spherical with a diameter of 45–60 nm. The nucleocapsid is surrounded by a lipid envelope. The genome consists of a single copy of a linear ssRNA that ranges in length from 12.7 to 15.7 kb. The host range of arteriviruses is limited, with horses, donkeys, swine, and monkeys infected by their respective species.

Single-stranded, negative-sense RNA viruses. The nucleocapsid of negative-sense RNA animal viruses contains an RNA-dependent RNA polymerase required for the transcription of the negative strand into the positive mRNAs. Virion RNA is neither capped nor polyadenylated. The group is divided into seven families: Orthomyxoviridae, Paramyxoviridae, Filoviridae, Rhabdoviridae, Bornaviridae, Arenaviridae, and Bunyaviridae. In addition, a genus known as *Deltavirus* found in this group is not assigned to a specific family.

Orthomyxoviridae. This group contains the genera *Influenzavirus A* (human influenza type A), *B*, and *C*, as well as the genus *Thogotovirus*. These are enveloped, spherical, pleomorphic particles 80–120 nm in diameter, with helical nucleocapsids. Their fragmented genome consists of 6–8 RNA

molecules, with genome sizes ranging from 10 to 14.6 kb. Surface projections for *Influenza A virus* are formed by the two major glycoproteins HA (hemagglutinin) and NA (neuraminidase). Serological classification by the World Health Organization (WHO) is based on these two neutralization antigens. *See* INFLUENZA VIRUS.

Paramyxoviridae. This group contains the subfamilies Paramyxovirinae and Pneumovirinae. The Paramyxovirinae include the genera *Respirovirus* (Sendai virus), *Rubulavirus* (mumps), and *Morbillivirus* (measles). The Pneumovirinae includes the genera *Pneumovirus* (respiratory syncytial virus) and *Metapneumovirus* (turkey rhinotracheitis virus). The enveloped, spherical particle, 150 nm in diameter, contains a single-stranded, nonfragmented RNA of approximately 15,000 nucleotides. *See* PARAMYXOVIRUS; RESPIRATORY SYNCYTIAL VIRUS; SENDAI VIRUS.

Filoviridae. This group includes two genera, the Marburg-like viruses and the Ebola-like viruses. Virions are pleomorphic but have mainly filamentous shape, sometimes with extensive branching. Length varies greatly between particles but the diameter is uniform (about 80 nm). Virions have a lipid membrane envelope derived from the host cell plasma membrane, with surface projections of viral glycoprotein of about 10 nm in length. The genome is a nonsegmented, negative-stranded, linear RNA molecule of about 19 kb in size. The natural reservoir of filoviruses is still unknown. Human disease is fatal in 25% of Marburg virus infections and in 50–90% of Ebola virus cases. *See* EBOLA VIRUS; MARBURG VIRUS; FILOVIRIDAE.

Rhabdoviridae. This group contains the genera *Vesiculovirus* (vesicular stomatitis Indiana virus), *Lyssavirus* (rabies), *Ephemerovirus* (bovine ephemeral fever virus), *Novirhabdovirus* (infectious hematopoietic necrosis virus), *Cytorhabdovirus* (lettuce necrotic yellows virus), and *Nucleorhabdovirus* (potato yellow dwarf virus). The particles are enveloped, bullet-shaped, and measure about 100–430 nm long by 45–100 nm in diameter. The outer surface of virions is covered with projections 5–10 nm long and about 3 nm in diameter. The nucleocapsid is 30–70 nm in diameter and exhibits helical symmetry. The genome consists of a single molecule of linear, negative-sense ssRNA about 11–15 kb in size. They contain five types of structural proteins: L, G, N, P, and M. Spikes are formed by G glycoprotein, and the M (matrix) protein occupies the space between the nucleocapsid and the envelope. The RNA is complexed with the N protein to form the helical nucleocapsid. Virus infectivity is rapidly inactivated at 56°C (133°F) or through exposure to lipid solvents.

Bornaviridae. The sole genus in this family is *Bornavirus* (Borna disease virus). The virions have a spherical morphology with a diameter of 90 ± 10 nm containing an internal core of 50–60 nm and a limiting outer membrane envelope that appears to be covered with projections approximately 7 nm long. The genome consists of a single molecule of a linear, nonsegmented negative-stranded RNA of 8.9 kb in size. Horses and sheep have been regarded as the main natural hosts of Borna disease virus, and the virus can cause a fatal neurological disease in these animals.

Arenaviridae. The sole genus in this family is *Arenavirus*, typified by the lymphocytic choriomeningitis virus. Its enveloped, spherical, pleomorphic particle is 50–300 nm in diameter. In section it exhibits granular appearance, due to ribosomes of host cell origin. A dense lipid envelope is covered by club-shaped projections 8–10 nm in length. The genome consists of two ssRNA molecules, L and S, of about 7.5 kb and 3.5 kb in size. Virus infectivity is inactivated by treatment with organic solvents or by exposure to ultraviolet (UV) and gamma irradiation. The reservoir hosts for this family are almost all rodents.

Bunyaviridae. This large family has been subdivided into five genera: *Bunyavirus* (bunyamwera), *Hantavirus* (Hantaan virus), *Phlebovirus* (Rift Valley fever virus), *Nairovirus* (Dugbe virus), and *Tospovirus* (tomato spotted wilt virus; Fig. 4). The virions are generally spherical or pleomorphic, 80–120 nm in diameter, and display surface glycoprotein projections of 5–10 nm embedded in a lipid bilayered envelope approximately 5 nm thick. The envelopes are usually derived from Golgi membranes in the host cell or sometimes cellular surface membranes. The genome contains three molecules of ssRNA, designated L (large), M (medium), and S (small), which total 11–19 kb. Viruses in the genera *Bunyavirus*, *Nairovirus*, and *Phlebovirus* are capable of alternately replicating in vertebrates and arthropods. Aerosol transmission occurs in hantaviral infections, while plant-to-plant transmission by thrips (tiny winged insects in the genus *Thysanoptera*) is necessary for Tospoviruses.

Deltavirus. These virions are roughly spherical with an average diameter of 36–43 nm. An outer envelope contains lipid and three envelope proteins obtained from the coinfecting helper *hepadnavirus*. An inner nucleocapsid of 19 nm comprises the RNA genome of hepatitis delta virus (HDV) and approximately 70 copies of the only HDV-encoded protein, known as delta antigen. The genome consists of a single molecule of circular negative-sense ssRNA about 1.7 kb in length. HDV requires the presence of a helper *hepadnavirus* to provide envelope proteins. It is therefore considered a subviral satellite virus. Natural HDV infection is found only in humans with hepatitis B virus as helper virus.

Double-stranded RNA viruses. The dsRNA vertebrate viruses contain two groups, the Reoviridae and the Birnaviridae.

Reoviridae. The *Reoviridae* include the nine genera *Orthoreovirus* (mammalian orthoreovirus), *Orbivirus* (bluetongue virus), *Rotavirus* (Rotavirus A), *Coltivirus* (Colorado tick fever virus), *Aquareovirus* (Aquareovirus A), *Cypovirus* (Cypovirus 1; see Fig. 3), *Fijivirus* (Fiji disease virus), *Phytoreovirus* (wound tumor virus in plants and insects), and *Oryzavirus* (rice ragged stunt virus). (See Fig. 4 for *Fjivirus*, *Phytoreovirus*, and *Oryzavirus*.) The

nonenveloped virions within this family have icosahedral symmetry but may appear spherical in shape. They each have a capsid, which is made up of concentric protein layers organized in one, two, or three distinct capsid shells. The overall diameter of the capsid is 60–80 nm. The nine genera may be divided into two groups: one group contains those viruses in which virus particles have relatively large spikes situated at the 12 vertices of the icosahedron; the second group includes those genera with relatively smooth or almost spherical virus particles lacking surface projections. Virions from some genera can leave infected cells by budding, acquiring a membrane envelope derived from the cellular membranes; however, this envelope appears to be transient in most cases. The virions contain a genome of 10, 11, or 12 linear segments of dsRNA. The properties of the *Reoviruses* vary according to the genus. *Rotavirus*, in particular, has become recognized as a frequent cause of human diarrhea. *See* ROTAVIRUS.

DNA and RNA reverse transcribing viruses. Vertebrate viruses in this category include the families Hepadnaviridae, discussed under the DNA virus category previously, and the Retroviridae.

The Retroviridae contains the genera *Alpharetrovirus* (avian leucosis virus), *Betaretrovirus* (mouse mammary tumor virus), *Gammaretrovirus* (murine leukemia virus), *Deltaretrovirus* (bovine leukemia virus), *Epsilonretrovirus* (Walleye dermal sarcoma virus), *Lentivirus* (human immunodeficiency virus or HIV), and *Spumavirus* (chimpanzee foamy virus). These virions are spherical, enveloped, and 80–100 nm in diameter. Surface projections made of glycoprotein are about 8 nm in length. The internal core contains the viral nucleocapsid, which is apparently spherical. The viral genome consists of a dimer of linear, positive-sense, ssRNA, each monomer 7–11 kb in length. There are two envelope proteins encoded by viral genes. Other proteins include a protease and a reverse transcriptase, which is essential to the process of producing a DNA copy of the RNA genome. This copy is then integrated into the host cell chromosome in a successful infection. Retroviruses have been implicated in human as well as animal leukemias, lymphomas, sarcomas, and immunodeficiencies, such as acquired immune deficiency syndrome (AIDS). *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS).

Families of Viruses Infecting Algae, Fungi, Yeast, and Protozoa

The shapes and sizes of characteristic members of the families of viruses that infect algae, fungi, yeast, and protozoa are presented in **Fig. 2**.

DNA viruses. Viruses in this group include one family, the Phycodnaviridae, and a genus known as *Rhizidiovirus* that is not assigned to a specific family.

Phycodnaviridae. Three genera include *Chlorovirus* (*Paramecium bursaria Chlorella virus 1*), *Prasinovirus* (*Micromonas pusilla virus SP1*), *Prymnesiovirus* (*Chrysochromulina brevifilum virus CbV-PW1*), and *Phaeovirus* (*Ectocarpus siliculosus virus 1*). These virions are nonenveloped, polyhedral particles with a shell surrounding an electron-dense core. They are 130–190 nm in diameter, with flexible, hairlike appendages extending from some of the vertices. The large genome consists of a dsDNA ranging from 160 to 380 kbp in size. The phycodnaviruses are ubiquitous in fresh water or seawater, and infect fresh-water and marine algae worldwide.

Rhizidiovirus. These virions are isometric, 60 nm in diameter. The genome consists of a single molecule of dsDNA. *Rhizidiomyces virus* infects the fungus *Rhizidiomyces*. The virus appears to be transmitted in a latent form in the zoospores of the fungus. No other viruses have been identified in this genus.

Double-stranded RNA viruses. RNA viruses infecting algae, fungi, yeast, and protozoa include dsRNA viruses, ssRNA viruses, and ssRNA viruses that use reverse transcriptase. The double-stranded RNA viruses include three families: Totiviridae, Partitiviridae, and Hypoviridae.

Totiviridae. Genera in this family include *Totivirus* (*Saccharomyces cerevisiae virus L-A*), *Giardiavirus* (*Giardia lamblia virus*), and *Leishmaniavirus* (*Leishmania RNA virus 1-1*). These virions have isometric symmetry and are 30–40 nm in diameter, nonenveloped, and lack surface projections. The genome consists of a single molecule of linear dsRNA, 4.6–7.0 kbp in size. These viruses are associated with latent infections of their fungal or protozoal hosts.

Partitiviridae. Genera in this family include *Partitivirus* (*Gaeumannomyces graminis virus*), *Chrysovirus* (*Penicillium chrysogenum virus*), *Alphacryptovirus* (White clover cryptic virus 1), and *Betacryptovirus* (White clover cryptic virus 2) [see Fig. 4 for *Alphacryptovirus* and *Betacryptovirus*]. These virions are isometric, nonenveloped, 30–40 nm in diameter, and contain two unrelated linear dsRNA segments (1.4–3.0 kbp in size). These viruses are associated with latent infections of their fungal and plant hosts.

Hypoviridae. The sole genus in this family is *Hypovirus* (*Cryphonectria hypovirus 1*). No true virions are associated with members of this family. Pleomorphic vesicles 50–80 nm in diameter, lacking any detectable viral structural proteins but containing dsRNA (9–13 kbp in size), are the only virus-associated particles that can be isolated from infected fungal tissue. Members infect the chestnut blight fungus, *Cryphonectria parasitica*.

Single-stranded RNA viruses. This group contains two families: Narnaviridae and Barnaviridae.

Narnaviridae. Two genera are found in this family: *Narnavirus* (*Saccharomyces cerevisiae narnavirus 20S RNA*) and *Mitovirus* (*Cryphonectria parasitica mitovirus 1-NB631*). No true virions are found associated with members of the genus *Narnavirus*. The *Narnavirus* genomes form ribonucleoprotein complexes containing linear single-stranded RNA of 2.5-kb and RNA-dependent RNA polymerases. These viruses infect the fungus *Saccharomyces cerevisiae*. The genus *Mitovirus* contains no reported virions. A single segment of dsRNA of 2.7 kb can be isolated from mitochondria of infected isolates. The virus

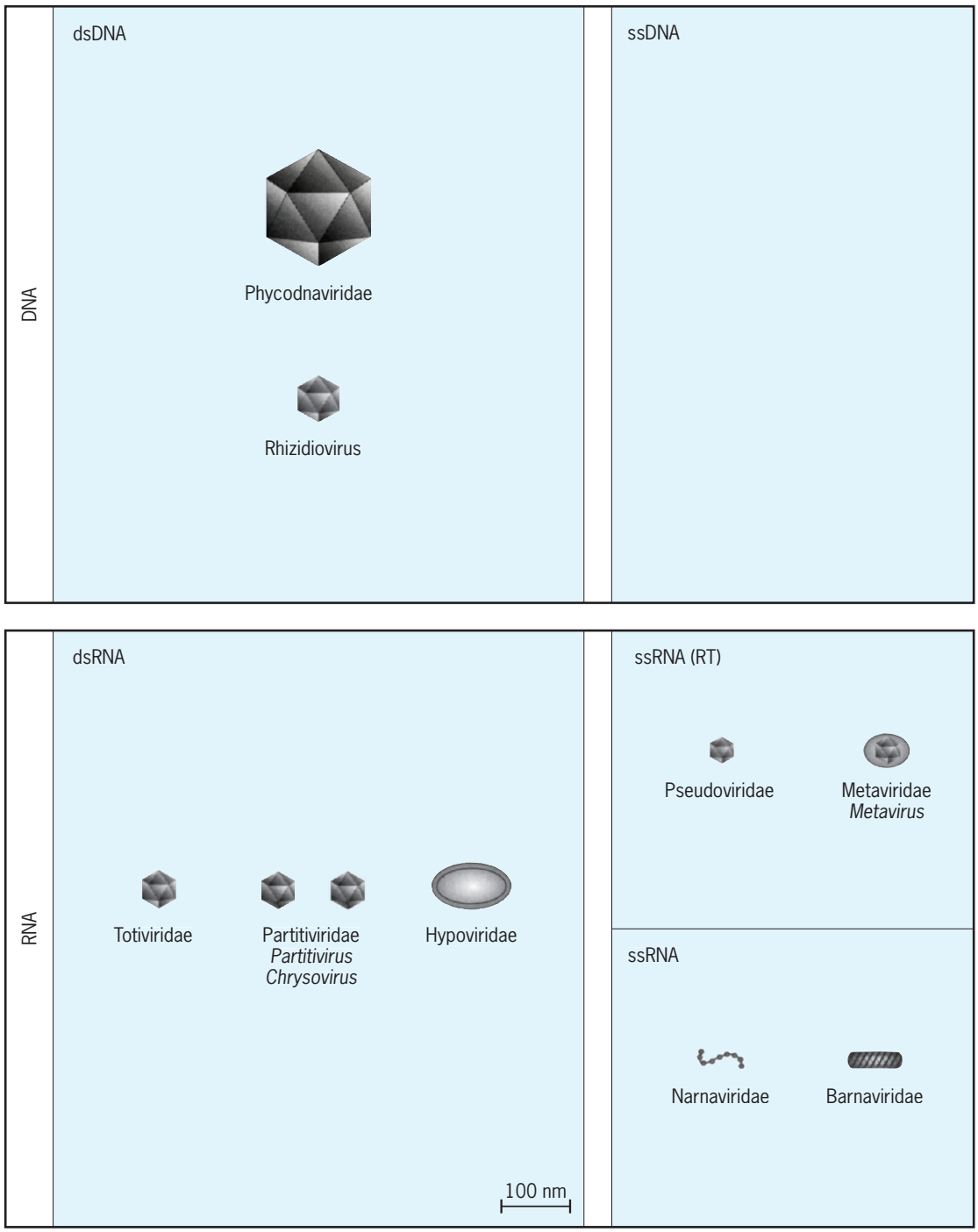


Fig. 2. Families and genera of viruses infecting algae, fungi, yeast, and protozoa. Individual frames separate taxa of viruses containing double-stranded and single-stranded genomes. Large boxes separate taxa of viruses having DNA or RNA genomes. All diagrams have been drawn approximately to the same scale to provide an indication of the relative sizes of the viruses. (*From M. H. V. van Regenmortel et al., eds., Virus Taxonomy: 7th Report of the International Committee on Taxonomy of Viruses, eds., p. 31, copyright 2000, with permission from Elsevier*)

infects members of the chestnut blight fungus, *Cryphonectria parasitica*.

Barnaviridae. The sole genus of this family is *Barnavirus* (mushroom bacilliform virus). Virions are nonenveloped, rod-shaped, and lack surface projections. Virions range 18–20 nm in width and 48–53 nm in length. The genome consists of a single linear molecule of a positive-sense ssRNA, 4.0 kb in size. The virus infects the common cultivated botton mushroom.

Single-stranded RNA viruses using reverse transcriptase. Viruses in this category include two families. *Pseudoviridae* and *Mataviridae*.

Pseudoviridae. Two genera are found in the family Pseudoviridae: *Pseudovirus* (*Saccharomyces cerevisiae Ty1 virus*) and *Hemivirus* (*Drosophila melanogaster copia virus*). The morphology of virus particles of this family is poorly known and variable. Both viruses make similar-looking particles, but those of *Saccharomyces cerevisiae Ty1*

virus are cytoplasmic whereas those of *Drosophila melanogaster copia virus* are nuclear. The major virion RNA molecule consists of a single strand of RNA of approximately 5–6 kb. A viral reverse transcriptase is used to copy the RNA and make a dsDNA "provirus," which is integrated into the host genome.

Metaviridae. Two genera are found in the family Metaviridae: *Metavirus* (*Saccharomyces cerevisiae Ty3 virus*) and *Errantivirus* (*Drosophila melanogaster gypsy virus*; **Fig. 3**). Morphology of particles is poorly characterized. Members include species which produce primarily or exclusively intracellular particles. Extracellular particles are enveloped with ovoid cores. Membranes are found in some members, apparently derived from the host cell. These viruses contain a reverse transcriptase as well as the positive strand RNAs.

Families and Genera of Viruses Infecting Invertebrates

The families of invertebrate viruses and the shapes and sizes of their characteristic members are presented in Fig. 3.

Double-stranded DNA viruses. These contain five families: Poxviridae, Baculoviridae, Iridoviridae, Polydnaviridae, and Ascoviridae.

Poxviridae. These viruses were previously described under vertebrate DNA viruses. However, the members of the subfamily Entomopoxvirinae strictly infect insects. Genera in this subfamily include *Entomopoxvirus A*, *B*, and *C*.

Baculoviridae. These are insect viruses. Two genera include *Nucleopolyhedrovirus* (*Autographa californica multiple nucleopolyhedrovirus*) and *Granulovirus* (*Cydia pomonella granulovirus*). The virions contain enveloped rod-shaped nucleocapsids. Nucleocapsid dimensions range 30–60 nm in diameter and 250–300 nm in length. The circular supercoiled dsDNA ranges 80–180 kbp in size.

Iridoviridae. These viruses were previously described under vertebrate DNA viruses. However, two genera of this family, *Iridovirus* and *Chloriridovirus*, strictly infect insects.

Polydnaviridae. These virions consist of nucleocapsids of uniform size (approximately 85 nm by 330 nm). A two-unit membrane envelope surrounds the ellipsoid nucleocapsid. Genomes consist of multiple dsDNA molecules of variable size, ranging from 2.0 to more than 31 kbp. The genera in this family are *Ichnovirus* (*Campoletis sonorensis ichnovirus*) and *Bracovirus* (*Cotesia melanoscela bracovirus*). These viruses have only been isolated from wasps; *Polydnaviruses* are unique among viruses in having obligate associations with parasitic wasps. The viral genome is inserted as a provirus into the host genome, from which new genome molecules are replicated.

Ascoviridae. These virions are bacilliform in shape, measuring about 130 nm in diameter by 200–400 nm in length. An envelope surrounds the inner particle. The sole genus in this family is *Ascovirus* (*Spodopter frugiperda ascovirus 1a*). These viruses cause disease in insects only, and are found primarily in lepidopterous larvae.

Single-stranded DNA viruses. This group includes only two families, the Parvoviridae and Circoviridae, which were previously described under vertebrate DNA viruses. The subfamily Densovirinae includes three genera responsible for infecting invertebrates: *Densovirus*, *Iteravirus*, and *Brevidensovirus*.

Double-stranded RNA viruses. These include two families: Reoviridae and Birnaviridae.

Reoviridae. These viruses were previously described under vertebrate RNA viruses. Only one genus in this family infects invertebrates: *Cypovirus* (*Cypovirus 1*). *Cypoviruses* have only been isolated from arthropods.

Birnaviridae. These viruses were previously described under vertebrate RNA viruses. Only one genus in this family infects invertebrates: *Entomobirnavirus* (*Drosophila X virus*).

Single-stranded, negative-sense RNA viruses. These include two families: Rhabdoviridae and Bunyaviridae.

Rhabdoviridae. These viruses were previously described under vertebrate RNA viruses. A large number of rhabdoviruses have not been assigned to an existing genus and there are a number of ungrouped viruses. An example is *Sigma virus*, which infects *Drosophila* species, conferring CO_2 sensitivity to infected insects.

Bunyaviridae. These viruses were previously described under vertebrate RNA viruses. They are capable of alternately replicating in vertebrates and arthropods, causing disease in the vertebrates while having little or no effect on the arthropods (ticks, mosquitoes, and other arthropod disease vectors).

Single-stranded, positive-sense RNA viruses. These include five families: Togaviridae, Flaviviridae, Picornaviridae, Tetraviridae, and Nodaviridae. They also contains the unassigned genus referred to as cricket paralysis like viruses.

Togaviridae. These viruses were previously described under vertebrate RNA viruses. The viruses in the genus *Alphavirus* are transmitted biologically between vertebrates by mosquitoes and other blood-feeding arthropods. Typically, infection in the arthropods becomes permanent.

Flaviviridae. These viruses were previously described under vertebrate RNA viruses. The viruses in the genus *Flavivirus* are usually transmitted biologically between vertebrates by mosquitoes and other blood-feeding arthropods.

Picornaviridae. These viruses were previously described under vertebrate RNA viruses. Transmission by arthropod vectors is not known, although *Encephalomyocarditis virus* has been isolated from mosquitoes and ticks.

Tetraviridae. Two genera are found in this family: *Betatetravirus* and *Omegatetravirus*. These virions are nonenveloped, roughly spherical, about 40 nm in diameter, and contain a single, positive-sense ssRNA segment of about 6.5 kb. All virus species are isolated from *Lepidoptera* species (moths and butterflies).

Nodaviridae. These viruses were previously described under vertebrate RNA viruses. All species of

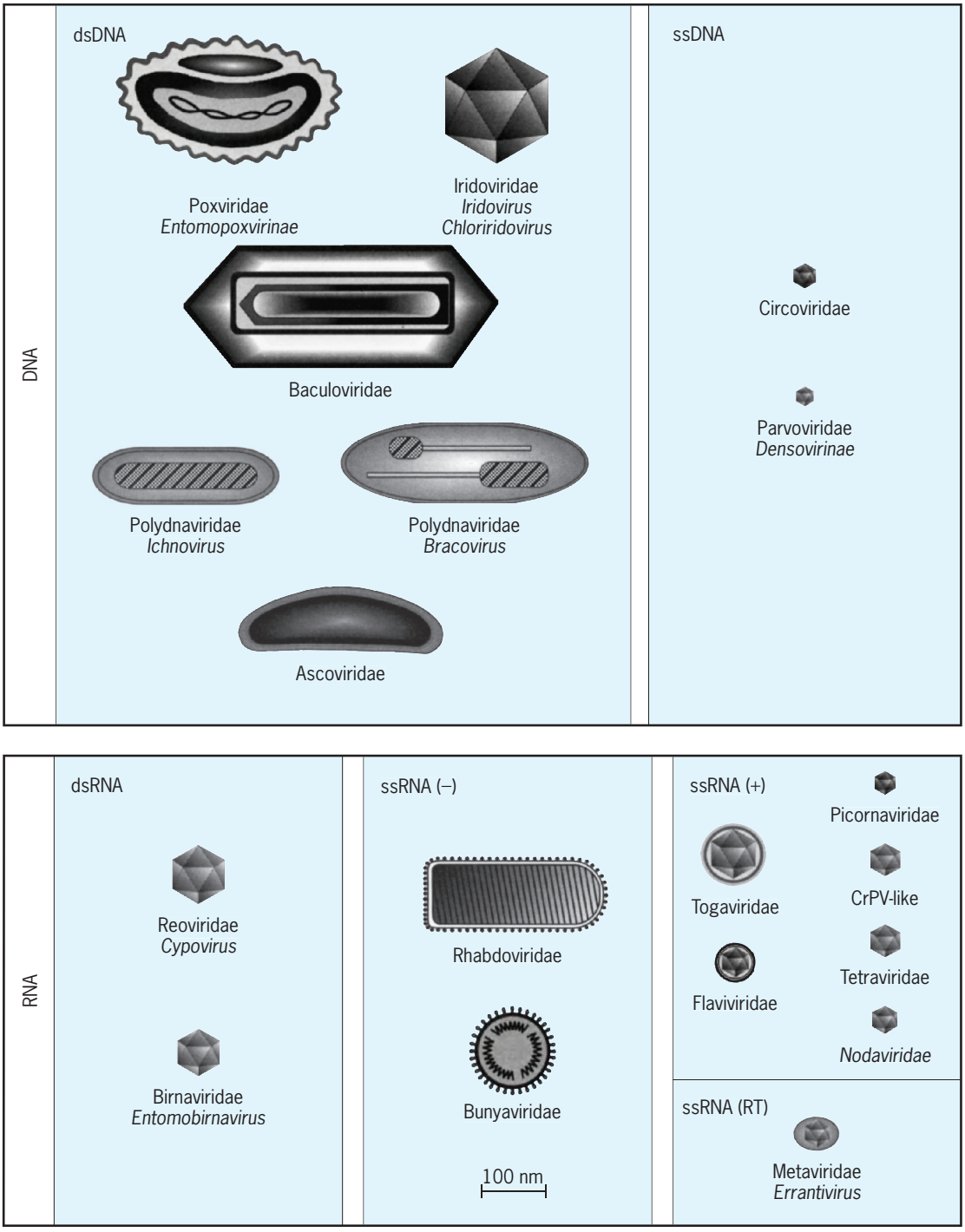


Fig. 3. Families and genera of viruses infecting invertebrates. Individual frames separate taxa of viruses containing double-stranded and single-stranded genomes. Large boxes separate taxa of viruses having DNA or RNA genomes. All diagrams have been drawn approximately to the same scale to provide an indication of the relative sizes of the viruses. (*From M. H. V. van Regenmortel et al., eds., Virus Taxonomy: 7th Report of the International Committee on Taxonomy of Viruses, p. 32, copyright 2000, with permission from Elsevier*)

the *Alphanodaviruses* were isolated in nature from insects.

Cricket paralysis-like viruses. This genus is not yet assigned to a family. These virions are roughly spherical with a diameter of 30 nm and no envelope. The virions exhibit icosahedral symmetry. The genome consists of a single molecule of linear, positive-sense ssRNA of 9–10 kb in size. All members of this genus have been isolated from invertebrate species.

Single-stranded RNA viruses using reverse transcriptase. This group contains only the family Metaviridae. These viruses were previously described under RNA viruses infecting algae, fungi, yeast, and protozoa. Members of the genus *Errantivirus* are known to infect species of *Drosophila*.

Families and Genera of Viruses Infecting Plants

The families of plant viruses and the shapes and sizes of their characteristic members are presented

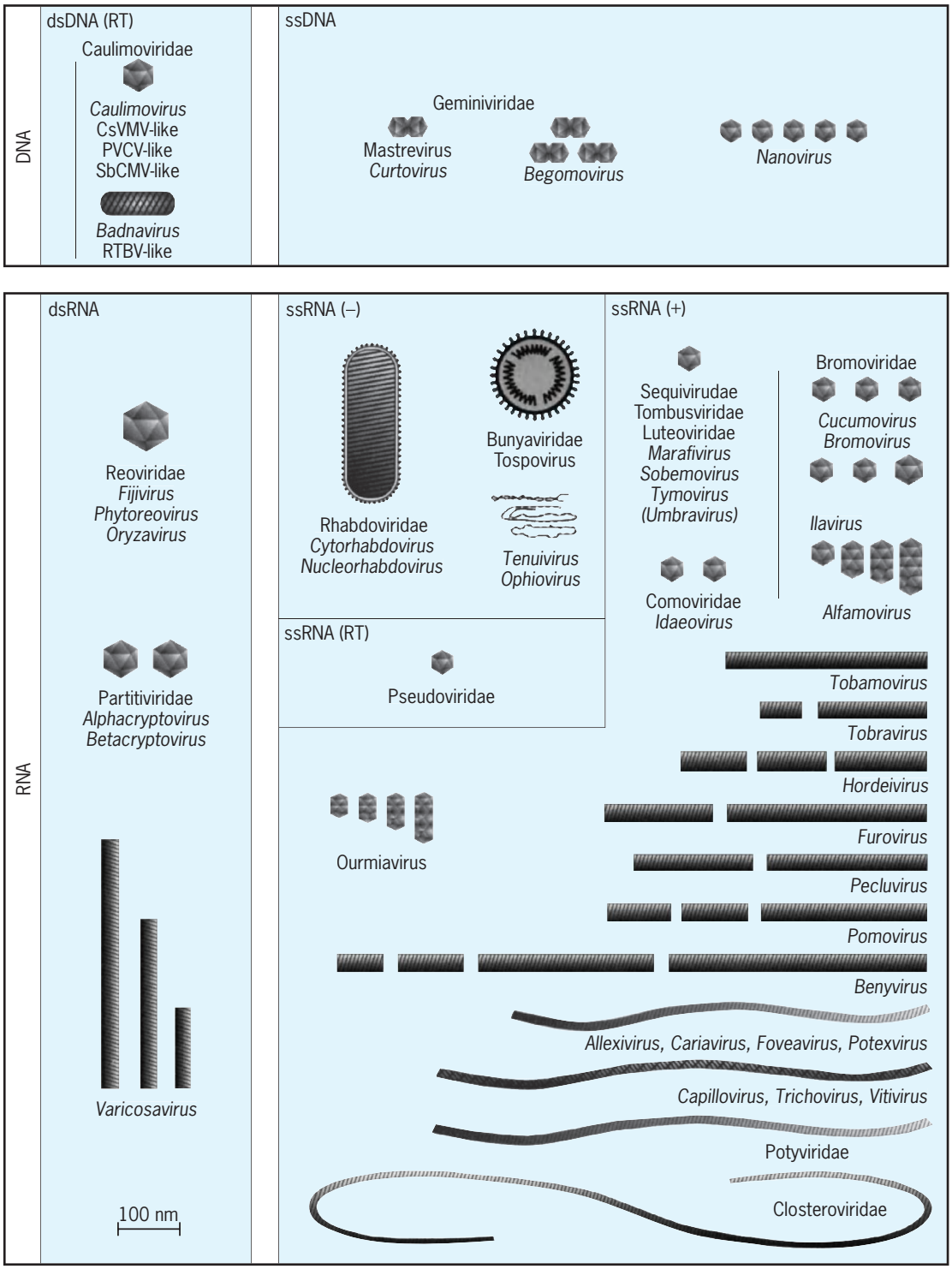


Fig. 4. Families and genera of viruses infecting plants. Individual frames separate taxa of viruses containing double-stranded and single-stranded genomes. Large boxes separate taxa of viruses having DNA or RNA genomes. All diagrams have been drawn approximately to the same scale to provide an indication of the relative sizes of the viruses. (*From M. H. V. van Regenmortel et al., eds., Virus Taxonomy: 7th Report of the International Committee on Taxonomy of Viruses, p. 33, copyright 2000, with permission from Elsevier*)

in **Fig. 4**. A number of plant viruses are placed into groups, rather than families. *See* PLANT VIRUSES AND VIROIDS.

Double-stranded DNA viruses. These include only the family Caulimoviridae. These virions are either bacilliform or isometric depending on the genus. There is no envelope. The genome consists of a single molecule of circular dsDNA of 7.2–8.1 kbp. Virions contain a reverse transcriptase. Genera in this family include *Caulimovirus*, Petunia vein clearing-like viruses, Soybean chlorotic mottle-like viruses, Cassava vein mosaic-like viruses, *Badnavirus*, and Rice tungro *bacilliform*-like viruses. The host ranges of most virus species are narrow. Many virus species are spread by vegetative propagation.

Single-stranded DNA viruses. These include one family, the Geminiviridae, and one unassigned genus called *Nanovirus*.

Geminiviridae. These virions apparently consist of two incomplete icosahedra with a total of 22 pentameric capsomers. The genome consists of a single molecule of circular ssDNA, 2.5–3.0 kb in size. Genera include *Mastrevirus*, *Curtovirus*, and *Begomovirus*. *Geminiviridae* are transmitted in nature by arthropods, in most cases by a single species.

Nanovirus. This genus is not yet assigned to a family. These virions are 17–20 nm in diameter, icosahedral, and lack envelopes. The genome is composed of several species of circular ssDNA approximately 1 kb in size. These viruses infect plants, including a range of leguminous species.

Double-stranded RNA viruses. These include two families, Reoviridae and Partitiviridae and the unassigned genus *Varicosavirus*.

Reoviridae. These viruses were previously described under vertebrate RNA viruses. Three genera in this family infect plants: *Fijivirus*, *Phytoreovirus*, and *Oryzavirus*. Viruses in these genera are typically transmitted between plants by arthropod vectors such as leafhoppers.

Partitiviridae. These viruses were previously described under RNA viruses infecting algae, fungi, yeast, and protozoa. Two genera are known to infect plants: *Alphacryptovirus* and *Betacryptovirus*. There are no known natural vectors for these viruses. The plant cryptoviruses are transmitted by ovule and by pollen to the seed embryo.

Varicosavirus. This genus is not yet assigned to a family. Virions are rod-shaped, about 320–360 nm in length, with a diameter of about 18 nm. The genome consists of dsRNA of two sizes, 7 kbp and 6.5 kbp. Viruses are transmitted by fungi of the genus *Olpidium* and are therefore soil- or waterborne. The type species for this genus is Lettuce big-vein virus.

Single-stranded negative-sense RNA viruses. These include two families, Rhabdoviridae and Bunyaviridae, plus two unassigned genera.

Rhabdoviridae. These viruses were previously described under vertebrate RNA viruses. Two genera in this family infect plants: *Cytorhabdovirus* and *Nucleorhabdovirus*. The viruses are primarily distinguished on the basis of the sites of virus maturation (cytoplasm: *Cytorhabdovirus*; nucleus: *Nucleorhabdovirus*).

Bunyaviridae. These viruses were previously described under vertebrate RNA viruses. One genus in this family is known to infect plants: *Tospovirus* (tomato spotted wilt virus). These viruses are known to be transmitted by thrips.

Tenuivirus. This genus is not yet assigned to a family. Virions have a thin filamentous shape, consisting of nucleocapsids 3–10 nm in diameter, with lengths proportional to the size of their RNA. The ssRNA genome consists of four or more segments, ranging in size from 11.9 to 9 kb. Plant hosts of tenuiviruses are all in the family Graminae.

Ophiovirus. This genus is not yet assigned to a family. Virions are naked filamentous nucleocapsids about 3 nm in diameter, with the shortest length about 760 nm. The genome consists of ssRNA of 11–12 kb in size. The natural hosts are dicotyledonous for two species, and tulip for the third species.

Single-stranded positive-sense RNA viruses. These include many families and unassigned genera, as follows.

Sequiviridae. Particles are isometric, about 30 nm in diameter. The genome consists of one molecule of infective, positive-sense RNA, 9–12 kb in size. Two genera are found in this family: *Sequivirus* (parsnip yellow fleck virus) and *Waikavirus* (rice tungro spherical virus). Transmission of these viruses is by aphids or leafhoppers; transmission is thought be dependent on a self-encoded helper protein.

Comoviridae. Virions are nonenveloped, 28–30 nm in diameter, and exhibit icosahedral symmetry. Genomes consist of two molecules of linear positive-sense ssRNA. Genera include *Comovirus* (cowpea mosaic virus), *Fabavirus* (broad bean wilt virus 1), and *Nepovirus* (tobacco ringspot virus). All members of the family have biological vectors: comoviruses are transmitted by beetles, fabaviruses by aphids, and many nepoviruses are transmitted by nematodes.

Potyviridae. Virions are flexible filaments (helical symmetry) with no envelope and are 11–15 nm in diameter. A single molecule of positive-sense ssRNA of 8.5–10 kb in size makes up the genome. Genera in this family include *Potyvirus* (potato virus Y), *Ipomovirus* (sweet potato mild mottle virus), *Macluravirus* (Maclura mosaic virus), *Rymovirus* (ryegrass mosaic virus), *Tritimovirus* (wheat streak mosaic virus), and *Bymovirus* (barley yellow mosaic virus). All members of the family *Potyviridae* form cytoplasmic cylindrical inclusion bodies during infection. Some members have a narrow host range, while most members infect an intermediate number of plants. Transmission to most hosts is accomplished by mechanical inoculation.

Sobemovirus. This genus is not yet assigned to a family. Virions are about 30 nm in diameter and exhibit icosahedral symmetry. The genome consists of a single molecule of positive-sense ssRNA, approximately 4.1–4.5 kb in size. The representative virus for this genus is southern bean mosaic virus. Sobemoviruses infect both monocotyledonous and dicotyledonous plants, but the natural host range of each species is narrow.

Marafivirus. This genus is not yet assigned to a family. These virions exhibit icosahedral symmetry, are 28–32 nm in diameter, do not have an envelope, and contain one molecule of linear, positive-sense ssRNA, about 6.5 kb in size. The representative virus for this genus is *Maize rayado fino virus*.

Luteoviridae. Virions are 25 to 30 nm in diameter, exhibit icosahedral symmetry, and have no envelope. The genome consists of a single molecule of linear, positive-sense ssRNA of about 5800 nts. Three genera are found in this family: *Luteovirus* (barley yellow dwarf virus), *Polerovirus* (potato leafroll virus), and *Enamovirus* (pea enation mosaic virus-1). Transmission is by specific aphid vectors.

Umbravirus. This genus is not yet assigned to a family. Umbraviruses do not form conventional virus

particles. Evidence suggests that the infective RNA is protected in lipid-containing structures. The representative virus is carrot mottle virus.

Tombusviridae. This is a very large family of plant viruses, with eight genera identified: *Aureusvirus*, *Avenavirus*, *Carmovirus*, *Dianthovirus*, *Machlomovirus*, *Necrovirus*, *Panicovirus*, and *Tombusvirus*. Virions exhibit an icosahedral symmetry with virions from the genera *Aureusvirus*, *Avenavirus*, *Carmovirus*, *Dianthovirus*, *Tombusvirus* having a rounded outline, a diameter of about 32–35 nm, and a granular surface due to the presence of 90 projections known as P domains. Virions in the genera *Machlomovirus*, *Necrovirus*, and *Panicovirus* lack the protruding domain and appear smooth. The genome consists of a single molecule of positive-sense, linear ssRNA ranging in size from 3.7 to 4.7 kb. The exceptions are the *Dianthovirus* virions, which contain two genomic RNAs. A characteristic virus representative is the *Tomato bushy stunt virus* (genus: *Tombusvirus*).

Tobamovirus. This genus is not yet assigned to a family. The representative for this genus is the tobacco mosaic virus, which is characterized by particles with a rigid helical rod, 18 nm in diameter and 300–310 nm in length. The genome consists of a single molecule of positive-sense linear ssRNA, 6.3–6.6 kb in size.

Tobravirus. This genus is not yet assigned to a family. Tobacco rattle virus is the best-known member of this group, which is characterized by two types of nonenveloped, rigid helical rods 21–23 nm in diameter either 180–215 nm or 46–115 nm in length, depending on the isolate. The genome consists of two molecules of linear positive-sense ssRNA; the first is 6.8 kb and the second ranges in size from 1.8 kb to about 4.5 kb.

Hordeivirus. This genus is not yet assigned to a family. Virions in this group are nonenveloped, elongated, and rigid; they are helically symmetric and about 20 nm in diameter and 110–150 nm in length. The genome contains three molecules of positive-sense ssRNA, each approximately 3–4 kb in size. The representative virus for this genus is the barley stripe mosaic virus.

Furovirus. This genus is not yet assigned to a family. Virions in this group are nonenveloped, hollow rods; they are helically symmetric and about 20 nm in diameter and 140–160 nm or 260–300 nm in length. The genome contains two molecules of linear, positive-sense ssRNA, with the first molecule about 6–7 kb and the second molecule about 3.5–3.6 kb in size. The representative virus for this genus is the soil-borne wheat mosaic virus.

Pomovirus. This genus is not yet assigned to a family. Virions in this group are nonenveloped, rod-shaped particles; they are helically symmetric and about 20 nm in diameter, with predominant lengths of 65–80, 150–160, and 290–310 nm. The genome contains three molecules of linear, positive-sense ssRNA, with the first molecule about 6 kb, the second about 3–3.5 kb, and the third 2.5–3 kb in size. The representative virus for this genus is the potato mop-top virus.

Pecluvirus. This genus is not yet assigned to a family. Virions in this group are nonenveloped, rod-shaped particles; they are helically symmetric and about 21 nm in diameter, with two predominant lengths of 190 and 245 nm. The genome contains two molecules of linear, positive-sense ssRNA, with the first molecule about 5900 nts and the second molecule about 4500 nts. The representative virus for this genus is the peanut clump virus.

Benyvirus. This genus is not yet assigned to a family. Virions in this group are nonenveloped, rod-shaped particles; they are helically symmetric and about 20 nm in diameter, with predominant lengths of about 85, 100, 265, and 390 nm. The genome contains four molecules of linear, positive-sense ssRNA, ranging in size from 1.3 kb to 6.7 kb. The representative virus for this genus is the beet necrotic yellow vein virus.

Bromoviridae. Virions in this group are either spherical with icosahedral symmetry (diameter 26–35 nm) or bacilliform (diameters of 18–26 nm and lengths from 30 to 85 nm). Genomes consist of three linear, positive-sense ssRNAs (total genome length is approximately 8 kb). Genera in this group include the *Alfamovirus* (alfalfa mosaic virus), *Bromovirus* (brome mosaic virus), *Cucumovirus* (cucumber mosaic virus), *Ilarvirus* (tobacco streak virus), and *Oleavirus* (olive latent virus 2). The natural host range of the viruses ranges from very narrow (genus *Bromovirus*) to extremely broad (genus *Cucumovirus*). Transmission predominantly occurs by insects, usually as a mechanical vector and not a persistent infection of the insect. Several viruses in this family are responsible for major disease epidemics in crop plants.

Ourmiavirus. This genus is not yet assigned to a family. Virions in this group constitute a series of particles with conical ends and cylindrical bodies 18 nm in diameter. There is no envelope. The genome contains three segments of linear, positive-sense ssRNA. The representative virus for this genus is the *Ourmia melon virus*.

Idaeovirus. This genus is not yet assigned to a family. The virions are isometric, about 33 nm in diameter, and not enveloped. The genome contains three segments of linear, positive-sense ssRNA of about 5.5 kb, 2.2 kb, and 1 kb in size. The representative virus for this genus is the raspberry bushy dwarf virus.

Closteroviridae. Virions in this group are very flexuous filaments with helical symmetry, about 12 nm in diameter, with the length varying according to the genus and/or the individual species. The genome consists of a single molecule of linear, positive-sense ssRNA ranging in size from 15.3 to 19.3 kb. Genera include *Closterovirus* (beet yellows virus) and *Crinivirus* (lettuce infectious yellows virus). The host ranges of individual virus species are restricted. Natural vectors include aphids and whiteflies.

Capillovirus. This genus is not yet assigned to a family. The virions are nonenveloped flexuous filaments with helical symmetry, about 12 nm in diameter and 640–700 nm in length. The genome contains linear, positive-sense ssRNA of 6.5–7.4 kb in size. The

representative virus for this genus is the apple stem grooving virus.

Trichovirus. This genus is not yet assigned to a family. These virions are nonenveloped flexuous filaments with helical symmetry, about 12 nm in diameter and 640–760 nm in length. The genome contains a single linear, positive-sense ssRNA of 7.5 kb in size. The representative virus for this genus is the apple chlorotic leaf spot virus.

Vitivirus. This genus is not yet assigned to a family. Virions are nonenveloped flexuous filaments with helical symmetry, about 12 nm in diameter and 725–825 nm in length. The genome contains a single linear, positive-sense ssRNA of 7.5 kb in size. The representative virus for this genus is the grapevine virus A virus.

Tymovirus. This genus is not yet assigned to a family. Virions are nonenveloped, icosahedral particles with a diameter of about 30 nm. The genome contains a single linear, positive-sense ssRNA of 6.3 kb in size. The representative virus for this genus is the turnip yellow mosaic virus.

Carlavirus. This genus is not yet assigned to a family. These virions are nonenveloped flexuous filaments with helical symmetry, 12–15 nm in diameter and 610–700 nm in length. The genome contains a single linear, positive-sense ssRNA of 7.4–7.7 kb in size. The representative virus for this genus is the carnation latent virus.

Potexvirus. This genus is not yet assigned to a family. These virions are nonenveloped flexuous filaments with helical symmetry, 13 nm in diameter and 470–580 nm in length. The genome contains a single linear, positive-sense ssRNA of 5.9–7.0 kb in size. The representative virus for this genus is the potato virus X.

Allexivirus. This genus is not yet assigned to a family. Virions in this group are nonenveloped highly flexible filaments, 12 nm in diameter and 800 nm in length. The genome contains a single linear, positive-sense ssRNA of 9.0 kb in size. The representative virus for this genus is the shallot virus X.

Foveavirus. This genus is not yet assigned to a family. These virions are nonenveloped flexuous filaments with helical symmetry, 12 nm in diameter and 800 nm in length. The genome contains a single linear, positive-sense ssRNA of 8.4–9.3 kb in size. The representative virus for this genus is the apple stem pitting virus.

Single-stranded RNA viruses using reverse transcriptase. This group contains only one family, the Pseudoviridae. These viruses were previously described under RNA viruses infecting algae, fungi, yeast, and protozoa. Members of this family are commonly referred to as LTR retrotransposons, and as such form an intrinsic and significant part of the genome of many plants.

Families and Genera of Viruses Infecting Bacteria

The families of bacteria, *Mycoplasma*, and Archeae viruses and the shapes and sizes of their characteristic members are presented in **Fig. 5**. Bacterial viruses are also known as bacteriophages or phages. They may be tailed or nontailed. *See* BACTERIOPHAGE.

Double-stranded DNA viruses. These include nine families and one unassigned genus.

Myoviridae. These virions have contractile tails with a central core surrounded by a helical contractile sheath. The head is elongated (or icosahedral), 110 by 80 nm, and the tail measures 113 by 16 nm. Particles contain at least 21 structural proteins, two or three inside the head. Many enzymatic activities are present. The tail has a collar, base plate, six fibers, six spikes, and a central tube. Fibers initiate attachment events. Genomes range in size 34–160 kbp. Genera include T4-like viruses, P1-like viruses, P2-like viruses, Mu-like viruses, SPO1-like viruses, and ϕ–H-like viruses (which infect Archeae). These viruses infect bacteria, with the exception of the ϕ–H-like viruses.

Siphoviridae. These virions have long, noncontractile, thin tails, 65–570 nm long and 7–10 nm in diameter, which are often flexible. Tails are built of stacked disks of six subunits. Genera in this family include λ-like viruses, T1-like viruses, T5-like viruses, L5-like viruses, c2-like viruses, and ψ M1-like viruses. Heads are primarily icosahedral, and genomes consist of dsDNA of varying sizes. These viruses infect bacterial species including *Enterobacteria*, *Mycobacterium*, *Lactococcus*, and the archaea species *Methanobacterium*.

Podoviridae. These virions have short, noncontractile tails about 20 × 8 nm in dimension. Genera include the T7-like viruses, P22-like viruses, and the ϕ29-like viruses. Heads are assembled first and tail subunits are added to the completed head. Phage heads are icosahedral. The genomes contain dsDNA of varying sizes. These viruses infect a wide number of bacterial species, including *Enterobacteria*, *Pseudomonas*, *Vibrio*, and *Caulobacter*.

Tectiviridae. Virus particles in this family are nonenveloped and icosahedral, 63 nm in diameter, and have 20-nm-long spikes at the 12 vertices. There are two shells, the outer rigid and 3 nm thick, while the inner is 6 nm thick, containing a lipoprotein vesicle. There are 15 virus-specific proteins associated with the lipoprotein vesicle. The genome consists of linear dsDNA of 14.7–15.7 kbp in size. During attachment to pili of gram-negative bacteria, following the ejection of nucleic acid, a transient tail of 60 nm length develops. The well- known *Enterobacteria phage PRDI* is in this group.

Corticoviridae. The virus particles are icosahedral and 60 nm in diameter, with brushlike spikes at the 12 vertices. The particle contains circular, dsDNA about 9 kbp in size. The virion also contains the enzyme transcriptase. A representative virus is *Alteromonas phage PM2.*

Plasmaviridae. This family of enveloped DNA phages includes the *Acholeplasma phage L2*. The particles are rounded, pleomorphic (50–125 nm in diameter), and show a small densely stained center. The double-stranded, circular DNA of the genome is 1.2 kbp in size. Virus is released by budding from *Acholeplasma* host cells, without killing the host.

Lipothrixviridae. These virions are rigid rods with a helical core, 410 nm long and 38 nm in diameter, with protrusions arising asymmetrically from both

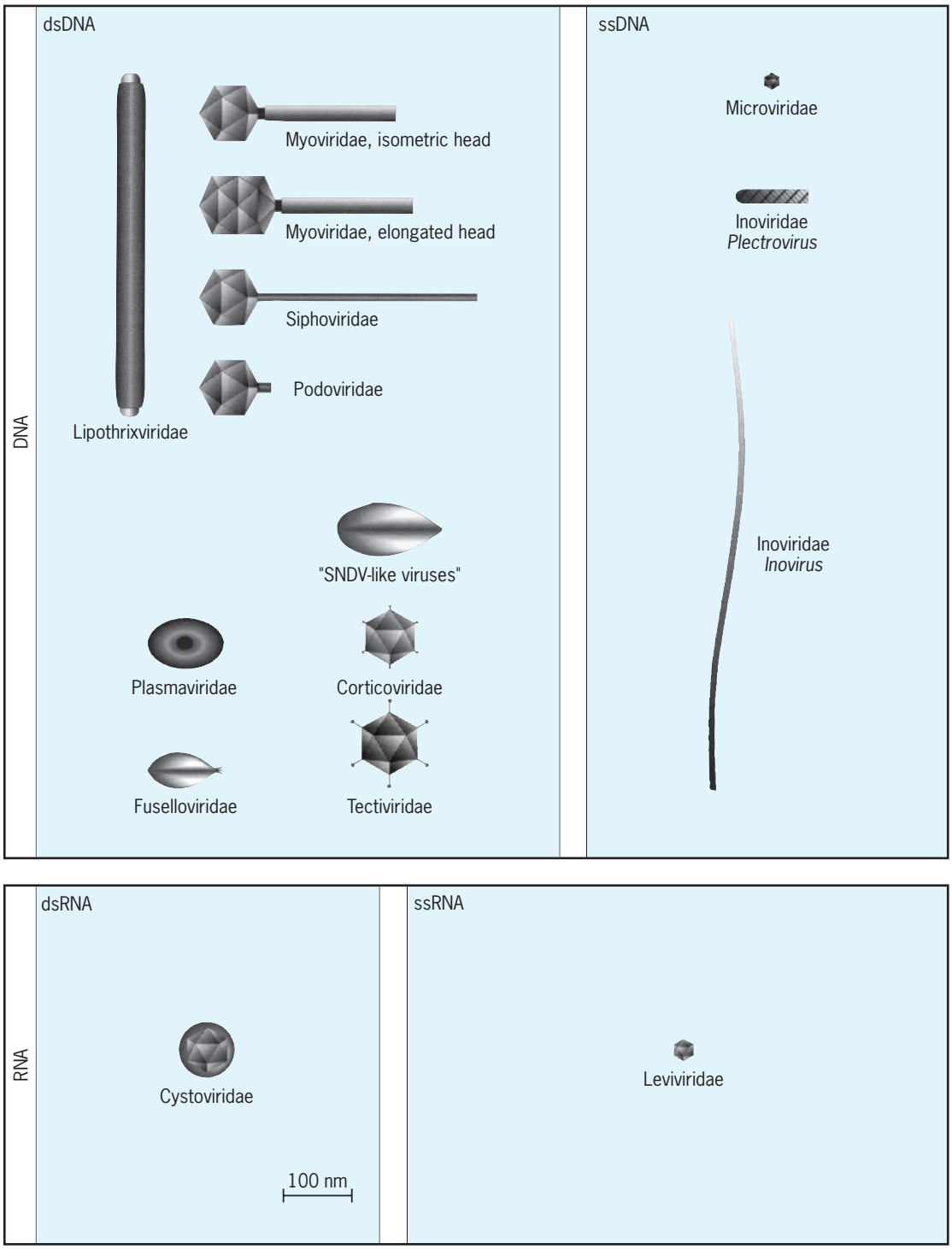


Fig. 5. Families and genera of viruses infecting bacteria. Individual frames separate taxa of viruses containing double-stranded and single-stranded genomes. Large boxes separate taxa of viruses having DNA or RNA genomes. All diagrams have been drawn approximately to the same scale to provide an indication of the relative sizes of the viruses. (*From M. H. V. van Regenmortel et al., eds., Virus Taxonomy: 7th Report of the International Committee on Taxonomy of Viruses, p. 34, copyright 2000, with permission from Elsevier*)

ends. The genome consists of one molecule of linear dsDNA of 15.9 kbp. A representative virus is *Thermoproteus virus 1*.

Rudiviridae. These virions are nonenveloped stiff rods, 23 nm in diameter by 500–780 nm in length. The genome is linear dsDNA, 33–36 kbp in size. A representative virus is *Sulfolobus virus SIRV-1*.

Fuselloviridae. These virions are lemon-shaped (60 by 100 nm), slightly flexible in appearance with short tail fibers attached to one pole. The genome contains circular, positively supercoiled dsDNA of 15.4 kbp in size. A representative virus is *Sulfolobus virus 1*.

Sulfolobus SNDV-like viruses. This genus is not yet assigned to a family. These virions are droplet-shaped, 80 by 180 nm. The genome is covalently closed circular dsDNA about 20 kbp in size. A representative virus is Sulfolobus virus SNDV.

Single-stranded DNA viruses. These include two families, Inoviridae and Microviridae.

Inoviridae. This family is characterized by nonenveloped rod-shaped phages which contain a single-stranded circular DNA genome within a cylindrical protein shell. Lengths vary from 700 nm to 2000 nm, although the diameter remains constant at about 7 nm. The genome consists of one molecule of infectious, circular positive-sense ssDNA. Genera in this family include *Inovirus* (*Enterobacteria phage M13*) and *Plectrovirus* (*Acholeplasma phage MV-L51*). Members of this family infect *Enterobacteria*, *Pseudomonas*, *Vibrio*, *Xanthomonas*, and *Acholeplasma*.

Microviridae. Virions exhibit icosahedral symmetry with projections at each of the 12 vertices. No envelope is present. Diameter of particles range from 26 to 32 nm. The genome for this family consists of one molecule of circular, positive-sense ssDNA, ranging in size from 4.4 to 6.0 kbp. Genera include *Microvirus*, *Spiromicrovirus*, *Bdellomicrovirus*, and *Chlamydiamicrovirus*. A representative virus for this family is *Enterobacteria phage ϕ X 174*. Microviridae infect *Enterobacteria*, *Bdellovibrio*, *Spiroplasma melliferum*, and *Chlamydia psittacii*.

Double-stranded RNA viruses. These include only the family Cystoviridae. The virions are enveloped, spiked bacteriophages about 85 nm in diameter. The envelope surrounds an icosahedral nucleocapsid (58 nm in diameter). The genome consists of three linear dsRNA segments of 6.4, 4.0, and 2.9 kbp in size. A representative virus is *Pseudomonas phage ϕ6*, which attaches to the pili of *Pseudomonas*.

Single-stranded RNA viruses. These include only the family Leviviridae. Virions are spherical and exhibit icosahedral symmetry (diameter of 26 nm). No envelope is present. The genome consists of one molecule of positive-sense ssRNA ranging in size from 3466 to 4276 nts. Two genera are found in this family: *Levivirus* (*Enterobacteria phage MS2*) and *Allolevivirus* (*Enterobacteria phage Qβ*). These viruses attach to pili of male *Enterobacteria*, *Caulobacter*, and *Pseudomonas*.

Marcia M. Pierce; M. E. Reichmann

Bibliography. D. M. Knight and P. M. Howley, (eds.), *Fundamental Virology*, 4th ed., Lippincott, Williams & Wilkins, Phildelphia, 2001; L. M. Prescott, J. P. Harley, and D. A. Klein, *Microbiology*, 6th ed., McGraw-Hill, New York, 2005; S. Schlesinger and M. Schlesinger, Viruses, in J. Lederberg (ed.), *Encyclopedia of Microbiology*, 2d ed., vol. 4, pp. 796–810, Academic Press, San Diego, 2000; M. H. V. Van Regenmortel et al. (eds.), *Virus Taxonomy: Seventh Report of the International Committee on Taxonomy of Viruses*, Academic Press, San Diego, 2000; B. A. A. Voyles, *The Biology of Viruses*, 2d ed., McGraw-Hill, Chicago, 2002.

Virus infection, latent, persistent, slow

Initially inapparent (covert) viral infections in which an equilibrium between the virus and the host has been established. Pathology may occur periodically, or chronically at a low level. If manifested in degenerative diseases of the central nervous system, these infections may result in death. The basis of these infections is an uncharacteristically low level of viral replication which may persist throughout the normal lifetime of the host, and which may follow the recovery from a more severe bout with the virus. The acute illness may have elicited production of antibodies, and an equilibrium between virus neutralization and antibody generation is established which may produce life-long immunity.

Persistent, slow, and latent infections are not easily differentiated. Persistent infections as described above are usually accompanied by detectable formation of antibodies without host pathology. (However, in the case of measles the persistent infection may change into a slow infection culminating in the degenerative brain disease termed subacute sclerosing panencephalitis, or SSPE.) Slow infections, while similar, eventually develop pathology. In latent infections, demonstrable presence of virus and symptoms of disease may be absent for a long time, with periodic outbreaks precipitated usually at a time when the immune system has been compromised in some way. This may occur through a variety of external influences, both physical and emotional.

Mechanisms. Several mechanisms are thought to account for the generation of latent, persistent, slow infections. Attenuation of the virulence of a virus can take place not only by immunological neutralization but also by spontaneous mutation of the virus itself. The most important group of mutants pertinent to these phenomena are the temperature-sensitive mutants, which replicate optimally at temperatures below the body temperature of the host, and very inefficiently at the elevated temperatures. The presence of temperature-sensitive mutants has been demonstrated in persistent infections of foot-and-mouth disease, coxsackie, sindbis, influenza, and equine encephalitis viruses. Another reason for diminished virulence in persistent infections may be due to the generation of defective interfering particles. Defective interfering particles are deletion mutants which are unable to replicate on their own. When they enter cells which are infected with the corresponding nonmutated virus they greatly reduce the number of infectious virus particles produced within these cells and thereby limit the extent of cell damage. The presence of defective interfering particles in persistent infections has been observed in the case of sendai, sindbis, lymphocytic choriomeningitis, and Newcastle disease viruses. Partial attenuation of viruses leading to persistent infections can also be brought about by host interferon production. *See* VIRULENCE; VIRUS, DEFECTIVE; VIRUS INTERFERENCE.

The integration of viral deoxyribonucleic acid (DNA) into the cellular genome in the case of DNA viruses, or reverse transcribed ribonucleic acid (RNA) retroviruses, often results in incomplete viral expression reminiscent of lysogeny in bacterial viruses. The afflicted cells may escape destruction by the virus itself as well as detection and destruction by the immune surveillance system. The visna

Slow virus infections*

Classification	Virus	Disease [host]
HUMANS		
RNA viruses		
Paramyxovirus	Measles variant	Subacute sclerosing panencephalitis (SSPE)
Rhabdovirus	Rabies	Rabies
Retrovirus	HTLV	T-cell leukemia, AIDS?
Picornavirus	Hepatitis A	Hepatitis
DNA viruses		
Papovavirus	JC, SV40-like	Progressive multifocal leukoencephalopathy (PML)
Unclassified	Hepatitis B	Hepatitis
ANIMALS		
RNA viruses		
Retrovirus	Visna	Meningoencephalitis [sheep]
	Maedi	Pneumonitis [sheep]
	Progressive pneumonia (PPV)	Pneumonitis [sheep]
	Equine infectious anemia	Hemolytic anemia, arteritis (equine infectious arteritis, or EIA) [horse]
	Gardner agent	Lower motor neuron [feral mouse]
	Xenotrophic C type	Hemolytic anemia, systemic lupus erythematosuslike [NZB mouse]
Arenavirus	Lymphocytic choriomeningitis (LCM)	Meningitis glomerulonephritis [mouse]
Paramyxovirus	Canine distemper	Encephalitis [dog]
Togavirus	Lactate dehydrogenase (LDV)	Elevated lactate dehydrogenase; mild nephritis [mouse]
Picornavirus	Theiler's agent	Demyelination [mouse]
DNA viruses		
Parvovirus	Aleutian disease (ADV)	Arteritis, anemia, nephritis [mink]
Papovavirus	SV40	Progressive multifocal leukoencephalopathy [monkey]
UNCONVENTIONAL AGENTS		
		Creutzfeldt-Jakob disease (CJD) [human]
		Kuru [human]
		Scrapie [sheep, goat]
		Transmissible mink encephalopathy

*After A. I. Braude, C. E. Davis, and J. Frerer (eds.), *Medical Microbiology and Infectious Diseases,* W. B. Saunders, 1981.

virus and possibly all the other retroviruses listed in the **table** belong to this group. In addition, most of the animal cancer-inducing retroviruses (not listed in the table) fall into this category, since even the nondefective retroviruses infect persistently and do not kill the host cell. *See* TUMOR VIRUSES.

Modified immunological tolerance can sometimes lead to the establishment of persistent infections. An example is lymphocytic choriomeningitis virus in mice, previously mentioned in connection with defective interfering particles. The virus can sometimes be contracted by the fetus while still in the womb, in which case a form of immunological tolerance develops with only a low level of antibodies generated at a later age. The circulating virus-antibody complexes which are at the root of the pathology of this disease are very slow to develop.

Sequestration of viruses in cells of either the peripheral or the central nervous systems probably plays an important role in the mechanism of latency. These primarily nondividing cells do not present a good environment for virus replication but can retain viruses for very long periods of time. Herpes virus, although not detected immediately, was shown to emerge when fragments of trigeminal ganglia were cultured in the laboratory. In intact organisms the virus is probably transported through the axoplasm in and out of these cells.

Slow infections. The term slow infection was introduced by Sigurdsson in 1954, while studying visna, maedi, and scrapie in Icelandic sheep. His criteria for a slow infection were as follows: a long preclinical period, extending for months or years from the time of first exposure; an extended period of a clinical course after appearance of symptoms; and confinement of pathology to a single organ or tissue system. Many of the slow infections in humans and other animals are represented in the table. A distinction is made between diseases in which a viral etiology was established and those in which the presence of virus particles cannot be demonstrated (unconventional agents). It now believed that the scrapie agent is a protein (mol wt 27,000–30,000) for which the name prion has been suggested. *See* PRION DISEASE; SCRAPIE.

Kuru, a neurologic affliction of people of the Fore and neighboring linguistic groups in New Guinea, was linked with mourning rites that included cannibalism, during which the population became infected by the Kuru agent, presumably through breaks in the skin or consumption of infected human tissue. A successful transmission of this disease to experimental animals prompted the investigation of other similar diseases, specifically the Creutzfeldt-Jakob disease, which has symptoms similar to Kuru. Eventually this rare but widespread

presenile dementia was successfully transmitted to chimpanzees. Since the discontinuation of the cannibalistic mourning rites among the Fore people, no new cases of Kuru have been observed in the subsequent generation of children.

Pathology. The tissue system most frequently affected in slow infections is the central nervous system. This is true of infections with papovaviruses (JC and other SV40-like viruses) and the unconventional agents Kuru, Creutzfeldt-Jakob disease, scrapie, and transmissible mink encephalopathy. The unconventional agents cause a degeneration that gives the tissue a spongy appearance when viewed under the light microscope; hence the name spongiform encephalopathies. The apparent sponginess is brought about by vacuolation of the neurons and the neuropil.

The papovaviruses cause demyelination through the destruction of the myelin sheath–furnishing cells, the oligodendrocytes. These symptoms, referred to as progressive multifocal leukoencephalopathy, have frequently occurred as a complication in persons with chronic lymphocytic leukemia, with Hodgkin's disease, or following treatment with immunosuppressants. Viruses, similar to JC (a human papovavirus), isolated from these cases are serologically distinct from other known papovaviruses, although they strongly cross-react with SV40.

Visna virus and subacute sclerosing panencephalitis (caused by an altered measles virus) are examples of inflammatory encephalitides. Lymphocytes and plasma cells infiltrate the central nervous system and form foci. The tissue becomes more firm; hence the name sclerosing.

The involvement of the central nervous system in slow infections manifests itself by disorientation, motor dysfunction, paralysis, and eventually death.

Latent infections. Herpes viruses are the subject of the most intensive studies of latency. The detection of infectious virus only after prolonged cultivation of explanted neural tissue implies the presence of the virus genome in a noninfectious state during latency. However, it can be argued that the infectious virus may be present in undetectably small quantities and the prolonged cultivation is necessary to increase its concentration to a detectable threshold. Despite the application of modern DNA technology to resolve this problem, the mechanism of establishment, maintenance, and reactivation of latent infections still remains obscure. *See* HERPES.

Human T-cell lymphotropic virus (HTLV) family. HTLV-I was isolated from several cases of cutaneous T-cell lymphomas in the United States. This virus is either identical or very closely related to the human adult T-cell leukemia virus (ATL), which is endemic in southwest Japan. A much rarer variant of the HTLV-I, the HTLV-II, has also been identified. HTLV-I (ATL) and HTLV-II seem to be the only human retroviruses positively identified in human malignancies. A third variant of this group, the HTLV-III or lymphadenopathy associated virus (LAV), was isolated from individuals suffering from acquired immune deficiency syndrome (AIDS). Antibodies against the last named virus have been found in 90–100% of AIDS patients, which indicates that they may have been infected with this agent. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS); ANIMAL VIRUS; VIRULENCE; VIRUS.

M. E. Reichmann

Bibliography. D. C. Bolton, M. P. McKinley, and S. B. Prusiner, Identification of a protein that purifies with the scrapie prion, *Science*, 218:1309–1310, 1982; H. Fraenkel-Conrat and R. R. Wagner (eds.), *Comprehensive Virology*, vol. 16, 1980; H. Hanfusa, M. E. Pullman, and A. Pinter (eds.), *Retrovirus and Disease*, 1989; J. A. Levy, H. Fraenkel-Conrat, and R. A. Owens., *Virology*, 3d ed., 1995; A. C. Minson, The state of the herpes genome, *Nature*, 304:477, 1983; R. Weiss, Retroviruses linked with AIDS, *Nature*, 300:12–13, 1984.

Virus interference

Inhibition of the replication of a virus by a previous infection with another virus. The two viruses may be unrelated, related, or identical. In some cases, virus interference may take place even if the first virus was inactivated. The term mutual exclusion has been applied to this phenomenon in bacterial viruses.

Several mechanisms of interference can be distinguished: (1) Inactivation of cell receptors by one virus may prevent subsequent adsorption and penetration by another virus. (2) The first virus may inhibit or modify cellular enzymes or proteins required for replication of the superinfecting virus. (3) The first virus may generate destructive enzymes or induce the cell to synthesize protective substances which prevent superinfection. (4) The first virus may generate defective interfering particles or mutants which may inhibit the replication of the infecting virus by competing with it for a protein (or enzyme) available in limited quantities; this type of viral interference has been called autointerference, and depends on a greater replicative efficiency of the defective interfering particles or mutants, compared to the infecting virus.

Examples of the first of these mechanisms are furnished by paramyxoviruses (parainfluenza), which contain in their envelope a neuraminidase enzyme. The enzyme hydrolyzes neuraminic (sialic) acid and its derivatives in the cell receptors utilized by the infecting virus or its progeny. Since closely related viruses of this group utilize the same receptors, superinfection cannot take place once the receptors have been destroyed.

Numerous examples of the second and third mechanisms are found among animal and bacterial viruses. The viral functions involved in modification of cellular proteins are usually also connected with the inhibition of normal cellular pathways. For example, poliovirus modifies the structure of cellular

ribosomes and renders them ineffective in cellular protein synthesis, while their participation in poliovirus protein synthesis is unimpaired. Superinfection with vesicular stomatitis virus (VSV) is abortive, because VSV protein synthesis is also inhibited. Similarly, ϕX 174 bacteriophage inhibits cellular deoxyribonucleic acid (DNA) synthesis as well as DNA synthesis of a superinfecting phage, and T4 phage generates enzymes which degrade DNA of the cell or of a superinfecting virus. Many animal viruses induce the synthesis of a group of host cell glycoproteins known as interferons. These interact with membranes of adjacent cells and confer on them resistance to viral infection. A variety of induced enzymatic activities results from this interaction between interferons and cells.

The generation of defective particles that interfere with the replication of the incoming virus has been reported for all major groups of animal viruses. This phenomenon may be important in persistent infection. *See* VIRUS, DEFECTIVE; VIRUS INFECTION, LATENT, PERSISTENT, SLOW.

No medical use of interference in prevention of viral infection has been made. The localized effect of the protecting virus and the requirement for its constant replication at relatively high levels reduces its efficacy as a protective agent. It is also difficult to find a protecting virus which is completely harmless to the organism.

Viral interference has been used in classifying relationships between paramyxoviruses, based on the use of identical receptors. It was also applied as a diagnostic test for rubella virus. Cells infected with this virus in tissue culture are protected from superinfection by other viruses (such as polio, measles, and influenza). The presence of rubella in the cells can be confirmed by the absence of cytopathic effects and cell death following superinfection by any of these viruses. *See* ANIMAL VIRUS; VIRULENCE; VIRUS.

M. E. Reichmann

Bibliography. H. Fraenkel-Conrat and R. R. Wagner (eds.), *Comprehensive Virology*, vol. 10, 1977; J. A. Levy, H. Fraenkel-Conrat, and R. A. Owens, *Virology*, 3d ed., 1995; S. Pestka, The purification and manufacture of human interferons, *Sci. Amer.*, 249:36–43, August 1983; M. E. Reichmann and W. M. Schnitzlein, Defective interfering particles of rhabdoviruses, *Curr. Topics Microbiol. Immunol.*, 86:123–168, 1979.

Viscosity

The material property that measures a fluid's resistance to flowing. For example, water flows from a tilted jar more quickly and easily than honey does. Honey is more viscous than water, so although gravity creates nearly the same stresses in honey and water, the more viscous fluid flows more slowly.

Shear viscosity. The viscosity can be measured in an experiment where the fluid of interest is sheared between two flat plates which are parallel to one another (**Fig. 1**). This is known as planar Couette flow. The shear flow created between the plates has the linear velocity profile given by Eq. (1), where v_x is

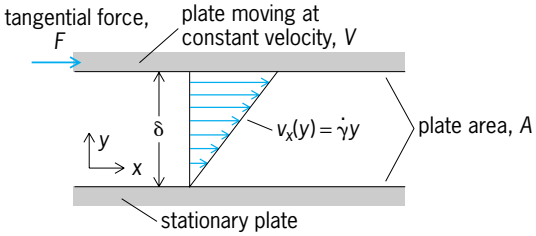


Fig. 1. Planar Couette flow. v_x = fluid velocity at distance y above the stationary plate, $\dot{\gamma}$ = velocity gradient or shear rate, δ = distance between plates.

$$v_x = \dot{\gamma}y \tag{1}$$

the velocity parallel to the plates at a perpendicular distance y above the stationary plate. The coefficient $\dot{\gamma}$, called the velocity gradient or shear rate, is given by V/δ, where δ is the distance between the plates. The shear stress is the ratio of the tangential force F needed to maintain the moving plate at a constant velocity V to the plate area A. It is expected that the shear stress increases with increasing shear rate but that the ratio of these two quantities depends only on the fluid between the plates. This ratio defines the shear viscosity, η, as in Eq. (2). The shear

$$\eta \equiv \frac{\text{shear stress}}{\text{shear rate}} = \frac{F/A}{V/\delta} \tag{2}$$

viscosity may depend on temperature, pressure, and shear rate.

Newton's law of viscosity. Isaac Newton is credited with first suggesting a model for the viscous property of fluids in 1687. Newton proposed that the resistance to flow caused by viscosity is proportional to the velocity at which the parts of the fluid are being separated from one another because of the flow. G. G. Stokes gave the first mathematical formulation of Newton's law of viscosity in 1845. Although Newton's law of viscosity is an empirical idealization, many fluids, such as low-molecular-weight liquids and dilute gases, are well characterized by it over a large range of conditions. However, many other fluids, such as polymer solution and melts, blood, ink, liquid crystals, and colloidal suspensions, are not described well by Newton's law. Such fluids are referred to as non-newtonian. *See* NON-NEWTONIAN FLUID; NON-NEWTONIAN FLUID FLOW.

For planar Couette flow (Fig. 1), the one-dimensional form of Newton's law of viscosity is given mathematically by Eq. (3), where τ_{yx} is the

$$\tau_{yx} = \mu \frac{dv_x}{dy} = \mu\dot{\gamma} \tag{3}$$

shear stress, and μ, a function of temperature and pressure, is the coefficient of viscosity or simply the viscosity. Therefore, by comparing Eqs. (2) and (3) the shear viscosity is equal to the coefficient of viscosity (that is, $\eta = \mu$) for a newtonian fluid.

TABLE 1. Viscosity conversions

Unit	poise	cp	Pa · s	$lb_m/(ft \cdot s)$	$lb_f \cdot s/ft^2$
1 poise* =	1	100	0.1	6.72×10^{-2}	2.089×10^{-3}
1 centipoise =	0.01	1	0.001	6.72×10^{-4}	2.089×10^{-5}
1 pascal-second† =	10	1000	1	0.672	2.089×10^{-2}
1 $lb_m/(ft \cdot s)$ =	14.88	1488	1.488	1	3.108×10^{-2}
1 $lb_f \cdot s/ft^2$ =	478.8	4.788×10^4	47.88	32.17	1

*1 poise = 1 dyne · s/cm^2 = 1 g/(cm · s).
†1 Pa · s = 1 kg/(m · s).

Because of this relation the shear viscosity is also often referred to as the viscosity. However, it should be clear that the two quantities are not equivalent; μ is a newtonian-model parameter, which varies only with temperature and pressure, while η is a more general material property which may vary nonlinearly with shear rate. *See* FLUID FLOW; NEWTONIAN FLUID.

Units and data. From Eqs. (2) and (3), the units of viscosity are given by force per area per inverse time. If in planar Couette flow, for example, 1 dyne of tangential force is applied for every 1 cm^2 area of plate to create a velocity gradient of 1 s^{-1}, then the fluid between the plates has a viscosity of 1 poise (=1 dyne · s/cm^2). Several viscosity units are in common use (**Table 1**). Comparison of the viscosities of different fluids (**Table 2**) demonstrates some general trends. For example, the viscosity of gases is generally much less than that of liquids. Whereas gases tend to become more viscous as temperature is increased, the opposite is true of liquids. Other data also show that increasing pressure tends to increase the viscosity of dense gases, but pressure has only a small effect on the viscosity of dilute gases and liquids. *See* DIMENSIONS (MECHANICS); UNITS OF MEASUREMENT.

Measurements. For newtonian fluids, there are several simple experiments from which the viscosity can be obtained. For example, when a solid sphere (for example, a ball bearing) of diameter D and density ρ_s falls through a fluid of viscosity μ and density ρ, the sphere eventually reaches a constant velocity, that is, the terminal velocity v_t. The flow around the sphere is well known if the Reynolds number of this flow, $\rho v_t D/\mu$, is roughly less than 0.1. Under these conditions the viscosity is given by Eq. (4), where g is

$$\mu = \frac{D^2(\rho_s - \rho)g}{18v_t} \tag{4}$$

the acceleration due to gravity. *See* CREEPING FLOW; REYNOLDS NUMBER.

For the flow of a newtonian fluid of density ρ and viscosity μ through straight conduit of circular cross section and radius R, such as a pipe or a capillary tube, the volume flow rate, Q, is proportional to the pressure drop per length, $\Delta p/L$. The viscosity is then given by the Hagen-Poiseuille law, Eq. (5). This relation assumes that the flow is laminar, which is valid when the Reynolds number for this flow, $2\rho Q/(\pi \mu R)$, is less than about 2100. *See* LAMINAR FLOW; PIPE FLOW.

$$\mu = \frac{\pi(\Delta p/L)R^4}{8Q} \tag{5}$$

The two methods of measuring the viscosity described above require assuming newtonian fluid behavior. The simple analysis used to derive Eqs. (4) and (5) fails for non-newtonian fluids. For this reason, devices have been developed to measure fluid properties in flows that closely approximate planar Couette flow (Fig. 1). One such flow can be created by shearing a fluid between a flat, circular plate and a cone (**Fig. 2**). The fluid is held in the gap between the cone and plate by the surface tension of the fluid. If

TABLE 2. Coefficients of viscosity of selected gases and liquids

Fluid	Temperature, °C (°F)	μ, centipoise
Air (gas)	−104 (−155)	0.01130
	0 (32)	0.01708
	74 (165)	0.02102
Hydrogen (gas)	0 (32)	0.00835
Ethane (gas)	0 (32)	0.00848
Water (liquid)	5 (41)	1.519
	20 (68)	1.002
	95 (203)	0.2975
Mercury (liquid)	0 (32)	1.685
Ethyl alcohol (liquid)	0 (32)	1.773
Olive oil (liquid)	10 (50)	138.0
Glycerin	0 (32)	12,110

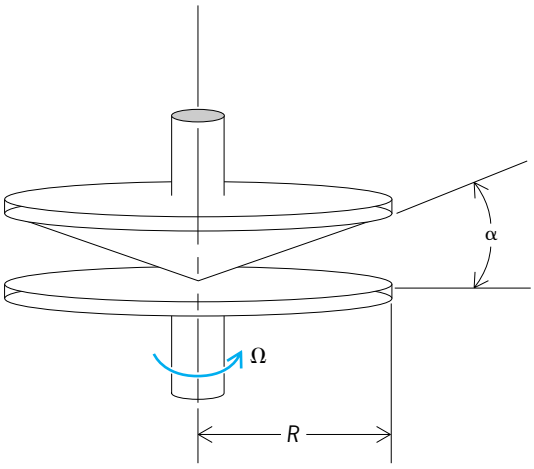


Fig. 2. Cone-and-plate geometry. α = cone angle, Ω = angular velocity of plate, R = radius of plate. (*After A. A. Collyer and P. W. Clegg, Rheological Measurement, Elsevier Applied Science, 1988*)

the cone angle α is small (roughly, less than 5°), then rotating either the plate (Fig. 2) or the cone at an angular velocity Ω generates a nearly constant shear rate of $\dot{\gamma} = \Omega/\alpha$ everywhere in the gap. The shear viscosity is proportional to the torque, $\mathscr{T}$, required to maintain steady rotation and is given by Eq. (6),

$$\eta(\dot{\gamma}) = \frac{3\mathscr{T}\alpha}{2\pi R^3 \Omega} \tag{6}$$

where R is the radius of the plate. By varying the angular velocity and measuring the applied torque, the cone-and-plate viscometer can be used to measure the shear viscosity over a large range of shear rates. Some sources of error in this measurement can arise from viscous heating, which can increase the temperature of the fluid; fluid inertia, which may cause the shear rate to vary in the gap; and nonideal geometry, which can create instabilities and nonideal flow.

Molecular origins. In classical fluid mechanics, the fluid is treated as a continuous medium without molecular structure. This approximation makes it necessary to use the concept of stress to account for the transport of momentum in the fluid caused by molecular interactions. It has been the goal of much research to derive expressions for the stress from molecular models of the fluid. One example is the kinetic theory of dilute gases.

Planar Couette flow (Fig. 1) can be considered for a dilute gas between two plates. The molecules of this gas can be modeled as nonattracting, hard spheres with diameter d and mass m. The motion of the spheres is chaotic with many collisions occurring between spheres themselves and between the spheres and the plates. The number density of spheres per volume n is small such that the mean free path is large (that is, the average distance traveled between collisions). From the kinetic theory of gases, the mean free path is

$$\lambda = 1/(\sqrt{2}\pi d^2 n)$$

and the average velocity of the gas molecules relative to the fluid velocity v is

$$\bar{c} = \sqrt{8k_B T/\pi m}$$

where k_B is Bolzmann's constant and T is the absolute temperature. The molecules exchange momentum during collisions; the slower molecules gain momentum, while the faster molecules lose momentum. For the flow described by Fig. 1, there is a flux of x momentum in the y direction which is the shear stress τ_{yx}. This model leads to a viscosity given by Eq. (7).

$$\mu = \frac{1}{3} nm\bar{c}\lambda = \frac{2}{3\pi^{3/2}} \frac{\sqrt{mk_B T}}{d^2} \tag{7}$$

See KINETIC THEORY OF MATTER; MEAN FREE PATH.

The simple hard-sphere model correctly predicts the independence of the viscosity on pressure and shear rate for very dilute gases. It also predicts that the viscosity increases as $\sqrt{T}$, which is qualitatively correct, although the viscosity of real gases depends more strongly on temperature. More refined models that employ potential energy functions lead to better quantitative agreement with data. *See* GAS; INTERMOLECULAR FORCES.

Although advanced kinetic theories of gases have been quite successful, a molecular theory of viscosity for liquids is necessarily more difficult. Whereas dilute gas molecules interact primarily in pairs as they collide, molecules in the liquid phase are in continuous interaction with many neighboring molecules. The concepts of average velocity and mean free path have little meaning for liquids. It is clear, however, that increasing temperature increases the mobility of molecules, thus allowing neighboring molecules to more easily overcome energy barriers and slip past one another. Such arguments lead to an exponential relation for the viscosity of the form given by Eq. (8),

$$\mu = A \exp\langle B/k_B T\rangle \tag{8}$$

where the constants A and B are characteristic of the liquid and must be determined experimentally. Although this is an approximate model, it does correctly give the decrease in viscosity with temperature that has been experimentally observed. *See* LIQUID.

Complex fluid behavior. Many non-newtonian fluids not only exhibit a viscosity which depends on shear rate (pseudoplastic or dilantant) but also exhibit elastic properties. These viscoelastic fluids require a large number of strain-rate-dependent material properties in addition to the shear viscosity to characterize them. The situation can become more complex when the material properties are time dependent (thixotropic or rheopectic). Fluids that are nonhomogeneous or nonisotropic require even more sophisticated analysis. The field of rheology attempts to deal with these complexities. *See* RHEOLOGY.

Lewis E. Wedgewood

Bibliography. R. B. Bird, W. E. Stewart, and E. N. Lightfoot, *Transport Phenomena*, 2d ed., Wiley, 2002; A. A. Collyer and D. W. Clegg (eds.), *Rheological Measurement*, 2d ed., Springer, 1998; D. R. Lide (ed.), *CRC Handbook Chemistry and Physics*, 86th ed., CRC Press, 2005; F. A. Morrison, *Understanding Rheology*, Oxford University Press, 2001; R. J. Silbey, R. A. Alberty, and M. G. Bawendi, *Physical Chemistry*, 4th ed., Wiley, 2004.

Vision

The sense of sight, which perceives the form, color, size, movement, and distance of objects. Of all the senses, vision provides the most detailed and extensive information about the environment. Conversely, blindness is recognized as more disabling than deafness or any other sensory handicap. In the higher animals, especially the birds and primates, the eyes and the visual areas of the central nervous system have developed a size and complexity far beyond the other sensory systems.

Visual stimuli. These are typically rays of light entering the eyes and forming images on the retina at

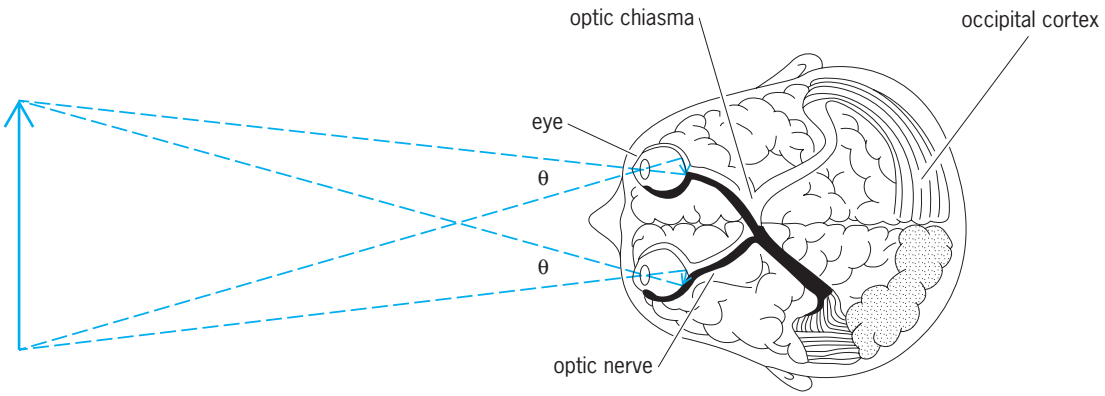


Fig. 1. Diagram showing the eyes and visual projection system. The visual angle θ is measured in degrees.

the back of the eyeball (**Fig. 1**). The intensity and wavelength characteristics of the light vary according to the light source and the object from which they are reflected. Human vision is most sensitive for light comprising the visible spectrum in the range 380–720 nanometers in wavelength. Sunlight and common sources of artificial light contain substantially all wavelengths in this range, but each source has a characteristic spectral energy distribution. In general, light stimuli can be measured by physical means with respect to their energy, dominant wavelength, and spectral purity. These three physical aspects of the light are closely related to the perceived brightness, hue, and saturation, respectively. *See* COLOR; LIGHT.

Atypical (sometimes called inadequate) stimuli for vision include momentary pressure on the eyeball, electric current through the eyes or head, a sudden blow on the back of the head, or disturbances of the central nervous system, caused by drugs, fatigue, or disease. Any of these may yield visual experiences not aroused by light. They are of interest because they show that the essentially visual character of the sensory experience is determined by the region stimulated (eyes, visual tracts) rather than by the nature of the stimulus. Indeed any observant person can detect swirling clouds or spots of "light" in total darkness or while looking at a homogeneous field such as a bright blue sky. These phenomena illustrate the spontaneous activity that is characteristic of the nervous system in general. They show that the visual system is continuously active, which in turn means that the effect of a stimulus is to modify existing activity and not merely to initiate new activity.

Anatomical basis for vision. The anatomical structures involved in vision include the eyes, optic nerves and tracts, optic thalamus, primary visual cortex, and higher visual areas of the brain. The eyes are motor organs as well as sensory; that is, each eye can turn directly toward an object to inspect it. The two eyes are coordinated in their inspection of objects, and they are able to converge for near objects and diverge for far ones. Each eye can also regulate the shape of its crystalline lens to focus the rays from the object and to form a sharp image on the retina. Furthermore, the eyes can regulate the amount of light reaching the sensitive cells on the retina by contracting and expanding the pupil of the iris. These motor responses of the eyes are examples of involuntary action that is controlled by various reflex pathways within the brain. *See* EYE (VERTEBRATE).

The process of seeing begins when light passes through the eye and is absorbed by the photoreceptors of the retina. These cells are activated by the light in such a way that electrical potentials are generated. These potentials are probably responsible for many features of the electroretinogram, an electrical response wave that can be detected by means of electrodes attached to the outside of the eye. Some of these potentials are the signs of electrochemical activity that serves to generate nerve responses in various successive neural cells (bipolars, ganglion cells, and others) in the vicinity of excitation. Finally, impulses emerge from the eye in the form of repetitive discharges in the fibers of the optic nerve. It must be emphasized, however, that the optic nerve impulses do not mirror exactly the excitation of the photoreceptors by light. Complex interactions within the retina serve to enhance certain responses and to suppress others. Furthermore it is a fact that each eye contains more than a hundred times as many photoreceptors as optic nerve fibers. Thus it would appear that much of the integrative action of the visual system has already occurred within the retina before the brain has had a chance to act.

The optic nerves from the two eyes traverse the optic chiasma. Figure 1 shows that the fibers from the inner (nasal) half of each retina cross over to the opposite side, while those from the outer (temporal) half do not cross over but remain on the same side. The effect of this arrangement is that the right visual field, which stimulates the left half of each retina, activates the left half of the thalamus and visual cortex. Conversely the left visual field affects the right half of the brain. This situation is therefore similar to that of other sensory and motor projection systems in which the left side of the body is represented by the right side of the brain and vice versa.

The visual cortex includes a projection area in the occipital lobe of each hemisphere. Here there appears to be a point-for-point correspondence between the retina of each eye and the cortex. Thus

the cortex contains a "map" or projection area, each point of which represents a point on the retina and therefore a point in visual space as seen by each eye. But the map is much too simple a model for cortical function. Vision tells much more than the location at which an object is seen. Visual tests show that other important features of an object such as its color, motion, orientation, and shape are simultaneously perceived. In monkey experiments, neurophysiologists have identified many types of cortical cells. each one responding selectively to these critical features. The two retinal maps are merged to form the cortical projection area. This merger allows the separate images from the two eyes to interact with each other in stereoscopic vision, binocular color mixture, and other phenomena. In addition to the projection areas on the right and left halves of the cortex, there are visual association areas and other brain regions that are involved in vision. Complex visual acts, such as form recognition, movement perception, and reading, are believed to depend on widespread cortical activity beyond that of the projection areas. *See* BRAIN.

Scotopic and Photopic Vision

Night animals, such as the cat or the owl, have eyes that are specialized for seeing with a minimum of light. This type of vision is called scotopic. Day animals such as the horned toad, ground squirrel, or pigeon have predominantly photopic vision. They require much more light for seeing, but their daytime vision is specialized for quick and accurate perception of fine details of color, form, and texture, and location of objects. Color vision, when it is present, is also a property of the photopic system. Human vision is duplex; humans are in the fortunate position of having both photopic and scotopic vision. Some of the chief characteristics of human scotopic and photopic vision are enumerated in the **table**.

Scotopic vision. This occurs when the rod receptors of the eye are stimulated by light. The outer limbs of the rods contain a photosensitive substance known as visual purple or rhodopsin. This substance is bleached away by the action of strong light so that the scotopic system is virtually blind in the daytime. Weak light causes little bleaching but generates neural inhibitory signals that lower the overall sensitivity of the eye. In darkness, however, the rhodopsin is regenerated by restorative reactions based on the transport of vitamin A to the retina by the blood. One experiences a temporary blindness upon walking indoors on a bright day, especially into a dark room or dimly lighted theater. As the eyes become accustomed to the dim light the scotopic system gradually begins to function. This process is known as dark adaptation. Complete dark adaptation is a slow process during which the rhodopsin is restored in the rod receptors of the retina. Faulty dark adaptation or night blindness is found in persons who lack rod receptors or have a dietary deficiency in vitamin A. These rare persons are unable to find their way about at night without the aid of strong artificial illumination. *See* VITAMIN A.

Dark adaptation is measured by an adaptometer, a device for presenting test flashes of light after various periods of time spent in the dark. The intensity of flash is varied to determine the momentary threshold for vision as dark adaptation proceeds. A 10,000-fold increase in sensitivity (that is, a reduction of 10,000 to 1 in the threshold intensity of flash) is often found to occur during a half-hour period of dark adaptation. By this time some of the rod receptors are so sensitive that only one elementary quantum (photon) of light is necessary to trigger each rod into action. A person can detect the presence of a flash of light that simultaneously affects only a few of the millions of rod receptors. Thus the scotopic sensitivity of the human eye approaches the ideal case of a receiver that is capable of responding to a single quantum of energy.

The variation of the scotopic threshold with wavelength of light is shown in the rod curve of **Fig. 2**. In spite of the variations in wavelength, the subject does not see any color when the intensity of the light is low enough to fall in the rod portion of the diagram. This scotopic vision is colorless or achromatic, in agreement with the saying that in the night all cats appear gray.

Normal photopic vision. Normal photopic vision has the characteristics enumerated in the table. Emphasis is placed on the fovea centralis, a small region at the very center of the retina of each eye.

Foveal vision. This is achieved by looking directly at objects in the daytime. In **Fig. 3** the image of a small object at F falls within a region almost exclusively populated by cone receptors. These are so closely packed together in the central fovea that their density is about 130,000 per square millimeter. Furthermore, each of the cones in the fovea is provided with

Characteristics of human vision

Characteristic	Scotopic vision	Photopic vision
Photochemical substance	Rhodopsin	Cone pigments
Receptor cells	Rods	Cones
Speed of adaptation	Slow (30 min or more)	Rapid (8 min or less)
Color discrimination	No	Yes
Region of retina	Periphery	Center
Spatial summation	Much	Little
Visual acuity	Low	High
Number of receptors per eye	120,000,000	7,000,000
Cortical representation	Small	Large
Spectral sensitivity peak	505 nm	555 nm

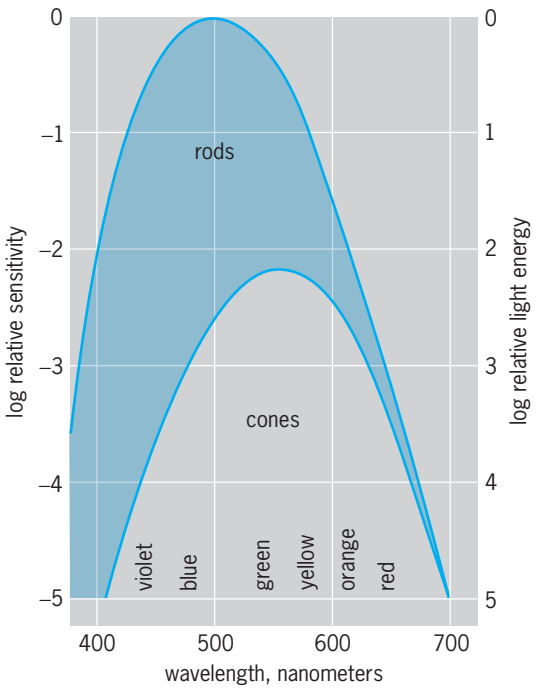


Fig. 2. Spectral sensitivity curves for human vision. The rod curve shows that scotopic vision, based on rhodopsin, is most sensitive to light of about 505 nm. The cone curve shows that photopic vision is generally less sensitive than scotopic vision, except for light at the red end of the spectrum.

a series of specialized nerve cells that process the incoming pattern of stimulation and convey it to the cortical projection area. In this way the cortex is supplied with superbly detailed information about any pattern of light that falls within the fovea centralis.

Peripheral vision. This is vision that takes place outside the fovea centralis (Fig. 3). As an example, look directly at a single letter at the center of a printed page. This letter, and a few letters immediately adjacent to it, appear clear and black because they are seen with foveal vision. The rest of the page is a blur in which the lines of print are seen as gray streaks. This is an example of peripheral vision. Vision actually extends out to more than 90° from center, so that one can detect moving objects approaching from either side. This extreme peripheral vision is comparable to night vision in that it is devoid of sharpness and color.

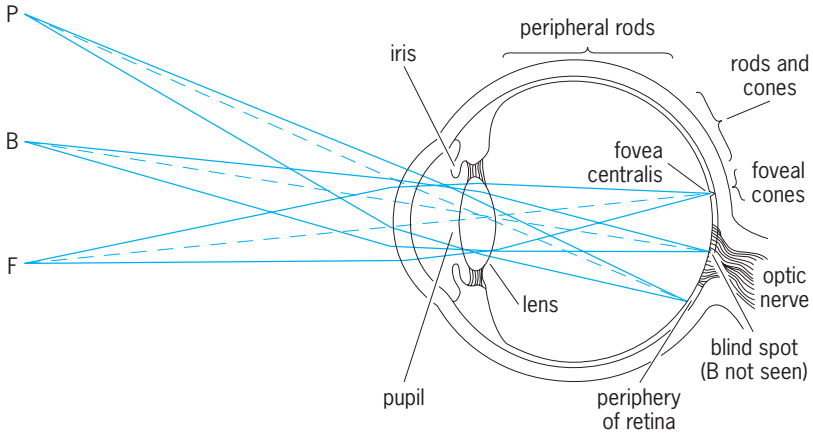


Fig. 3. Foveal and peripheral vision. Looking directly at F, the eye focuses the light from F on the central fovea, the region of clearest vision. P is seen poorly, and B not at all.

There is a simple anatomical explanation for the clarity of foveal vision as compared with peripheral vision. The cone receptors become less and less numerous in the retinal zones that are more and more remote from the fovea. In the extreme periphery there are scarcely any cones, and even the rod receptors are more sparsely distributed. Furthermore, the plentiful neural connections from the foveal cones are replaced in the periphery by network connections in which hundreds of receptors may activate a single optic nerve fiber. This mass action is favorable for the detection of large or dim stimuli in the periphery or at night, but it is unfavorable for visual acuity or color vision, both of which require the brain to differentiate between signals arriving from closely adjacent cone receptors.

Visual acuity. Visual acuity is defined as the ability to see fine details of an object. In Fig. 1 the small arrow at the back of each eye shows the image of the test object that is focused on the retina. Standard visual acuity is defined as the ability to see an object so small that the angle θ subtended at the eye is only 1 minute of arc, or 1/60 of a degree. At 20 ft (0.6 m) the size of such a test object is therefore only about 0.07 in., or 1.75 mm. The image of the object on the fovea (neglecting diffraction and optical aberrations that deteriorate the image) has a length of only 0.005 mm. Small as this is, it is twice the diameter of the smallest foveal cone receptor. One therefore comes to the conclusion that normal visual acuity approaches the limit imposed by diffraction and by the optical aberrations in the eye.

The specific forms of test object used for determining visual acuity yield different results. The angle θ in Fig. 1 can be made so small that it represents the size of test object that can barely be seen by the normal eye. This angular size can be infinitely small in the case of a bright point seen against a dark background. The stars at night provide a good example of this, since they are so distant that their angular subtense at the eye is practically zero. A long dark line can be detected against a bright field (for example a flagpole seen against the sky) if it subtends 1 second of arc (1/60 of a minute) at the eye. On the other hand, letters of the alphabet, as used in optometric wall charts, need to be composed of black lines 1 minute thick in order to be recognizable to the normal eye. A similar value holds for the lines of a grating (with parallel black and white lines of equal width) when viewed under good illumination.

The apparent discrepancy between acuity for single points or lines on the one hand, and for more complex forms on the other hand, can be explained by the effects of optical diffraction. This is a phenomenon resulting from the wave nature of light. It means that no optical image can ever be completely sharp and clear. The retinal image of a star is not a point but a blurred circle of light. The diameter of this blurred image is never less than 1 minute of arc,

no matter how small an angle the star itself may subtend. Thus the star is seen, provided that the blurred image is noticeably brighter than the surrounding field. In the case of a grating (black and white stripe pattern) the image of each line is blurred also. If the lines are too close (less than 1 minute) together, the blurred image of one overlaps the blurred image of the next and the separate lines cannot be resolved by the eye. One comes to the conclusion, then, that the chief factors limiting the visual acuity of the fovea are (1) optical diffraction, (2) the ability to discriminate relative brightnesses within the images blurred by diffraction, and (3) the compactness of the pattern of cone receptors.

Space and Time Perception

Spatial and temporal effects are clearly apparent in the sense of sight. These two effects enable the individual to be oriented with regard to space and time in the surrounding world, especially in the perception of motion and distance. *See* PERCEPTION.

Space perception. Elementary forms of space perception are vernier and stereoscopic discrimination. Here, the eye is required to judge the relative position of one object in relation to another (**Fig. 4**). The left eye, for example, sees the lower line as displaced slightly to the right of the upper. This is known as vernier discrimination. The eye is able to distinguish fantastically small displacements of this kind, a few seconds of arc under favorable conditions. If the right eye is presented with similar lines that are oppositely displaced, then the images for the two eyes appear fused into one and the subject sees the lower line as nearer than the upper. This is the principle of the stereoscope. Again it is true that displacements of a few seconds of arc are clearly seen, this time as changes in distance. The distance judgment is made not at the level of the retina but at the cortex where the spatial patterns from the separate eyes are fused together. The fineness of vernier and stereoscopic discrimination transcends that of the retinal mosaic and suggests that some averaging mechanism must be operating in space or time or both. Furthermore, experiments with random dot stereograms demonstrate that the averaging mechanism can produce an impression of depth without visible contours in the retinal images of the two eyes.

The spatial aspects of the visual field are also of interest. As has been previously indicated, good acuity is restricted to a narrowly defined region populated by densely packed cones at the center of the visual field. A somewhat larger central region, in which the cones are somewhat more sparsely distributed, is capable of good color vision. Farther out, however, vision mainly is mediated by the rod receptors; color and form vision become extremely poor. In these peripheral regions area and intensity are reciprocally related for all small sizes of stimulus field. A stimulus patch of unit area, for example, looks the same as a patch of twice the same area and half the luminance. This high degree of areal summation is achieved by the convergence of hundreds of rod receptors upon each single optic nerve fiber. It is the basis for the ability of the dark-adapted eye to detect large objects even on a dark night.

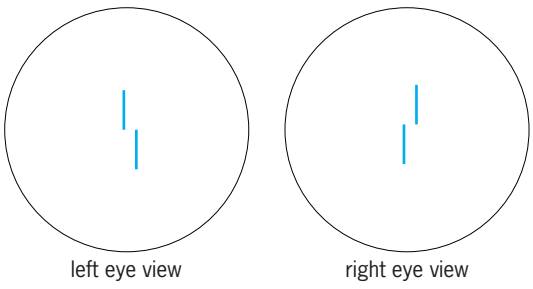


Fig. 4. Vernier and stereoscopic discriminations of space.

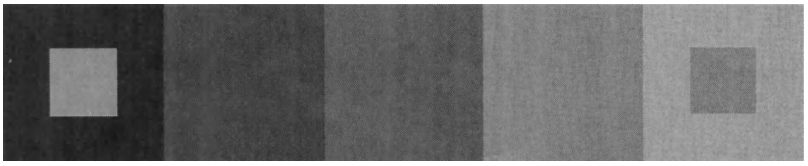


Fig. 5. Simultaneous contrast. The two small squares are physically equal, and each of the five strips is uniform. Spatial inhibitory effects cause the right square to look darker than the left square and cause the apparent darkening of the portion of each strip that is adjacent to a lighter strip.

In daytime vision, spatial inhibition, rather than summation, is most noticeable. The phenomenon of simultaneous contrast is present at a border between fields of different color or luminance. Thus a small square of gray paper appears darker on a light background than on a dark one; on a yellow background it appears bluish. The effect may originate in neural mechanisms of inhibition, such that stimulating a given region of a retina depresses the activity of regions immediately adjacent to it. This has the obvious effect of heightening contours and making forms more noticeable against their background (**Fig. 5**).

In photopic vision out of doors the eyes become light-adapted to an extent determined by such things as the time of day or the cloud cover. The luminance range of objects seen on a given day is typically no more than about 100 to 1 (two logarithmic units). Over this dynamic range the eyes appear to be capable of good spatial and temporal resolution, without the necessity for slowly adapting to the various levels of luminance.

Time perception. The temporal characteristics of vision are revealed by studying the responses of the eye to various temporal patterns of stimulation. When a light is first turned on, there is a vigorous burst of nerve impulses that travel from the eye to the brain. Continued illumination results in fewer and fewer impulses as the eye adapts itself to the given level of illumination. Turning the light off elicits another strong neural response. Afterimages are often seen at this time. A positive afterimage, resembling the original stimulus, is sometimes seen during the first fraction of a second after the light goes off. This is usually followed by a longer-lasting negative afterimage in which the color of the original object appears to be reversed. A clear afterimage may be produced by staring fixedly at the stimulus for at least

a half minute, then turning away and "projecting" the afterimage against a white or gray screen. The retinal size of the afterimage remains constant, so that its size in inches on a screen is proportional to the distance of the screen from the observer (Emmert's law). The afterimage is thought to arise chiefly from the fact that the affected region of the retina has been bleached, or otherwise changed photochemically, in such a way that its responses are different from those of the regions not stimulated. To some extent, however, visual aftereffects are due to central, rather than retinal, processes. *See* COLOR VISION.

The strength of a visual stimulus depends upon its duration as well as its intensity. Below a certain critical duration, the product of duration and intensity is found to be constant for threshold stimulation. A flash of light lasting only a few milliseconds may stimulate the eye quite strongly, providing its luminance is sufficiently high. A light of twice of the original duration will be as detectable as the first if it is given half the original luminance. However, even though two lights of the same energy may be equally detectable, a person may still be able to distinguish between them.

Voluntary eye movements enable the eyes to roam over the surface of an object of inspection. In reading, for example, the eyes typically make four to seven fixational pauses along each line of print, with short jerky motions between pauses. An individual's vision typically takes place during the pauses, so that one's awareness of the whole object is the result of integrating these separate impressions over time.

A flickering light is one that is going on and off (or undergoing lesser changes in intensity) as a function of time. At a sufficiently high flash rate (called the critical frequency of fusion, cff) the eye fails to detect the flicker, and the light pulses seem to fuse to form a steady light that cannot be distinguished from a continuous light that has the same total energy per unit of time. As the flash rate is reduced below the cff, flicker becomes noticeable, and at very low rates the light may appear more conspicuous than flashes occurring at higher frequency. The cff is often used clinically to indicate a person's visual function as influenced by drugs, fatigue, or disease. *See* PSYCHOLOGY.

Contrast sensitivity functions. Another approach to the study of vision is based on the concept of modulation, a term that has long been associated with periodic temporal changes in radio signals or in sound. But vision is particularly concerned, not with temporal variation, but with the spatial variations of light that characterize a pattern. The approach is to analyze pattern vision in terms of the contrast, or light modulation, that is measured within it. Furthermore, just as the methods of Fourier analysis have already yielded a good understanding of information transfer in radio transmission, the same methods are being applied with great success to pattern vision.

In Fourier terms, the simplest form of modulation is sinusoidal. Accordingly, an elementary form of visual pattern is the one in **Fig. 6**, where spatial frequency is 4 cycles per degree of visual angle (Fig. 1), average light intensity is I_{av}, and contrast (C) is $(I_{max} - I_{min})/(I_{max} + I_{min})$.

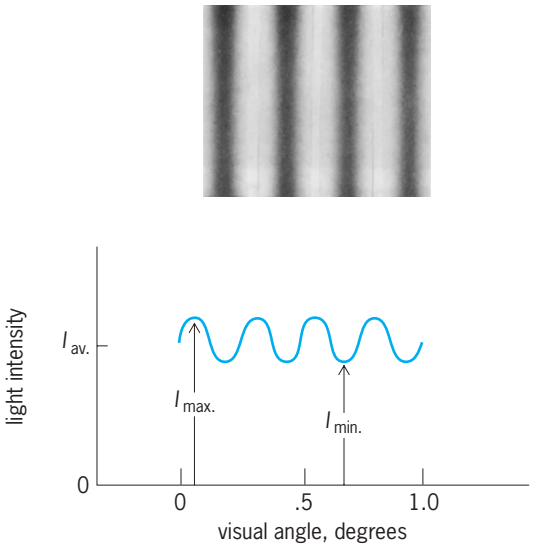


Fig. 6. Example of sinusoidal grating.

Sinusoidal gratings can conveniently be produced and controlled electronically on the face of a laboratory oscilloscope. To measure human contrast sensitivity, the contrast (C) of the grating is gradually reduced to threshold (C_0), that is, to a point where the grating is barely distinguishable from a uniform field of the same average intensity. Contrast sensitivity is defined as the reciprocal of this threshold value of contrast.

Several important facts about vision emerge from the sinusoidal grating experiments. (1) the entire range of spatial sensitivity can be explored in this way. A typical contrast sensitivity function for human vision (**Fig. 7**) peaks at about 3 to 4 cy/degree and becomes severely attenuated below 0.1 cy/deg and above 60 cy/deg. (2) prolonged viewing of a sinusoidal grating at high contrast causes a loss of contrast sensitivity for a grating of the same spatial frequency, but not for gratings of considerably higher or lower spatial frequency. This means that there

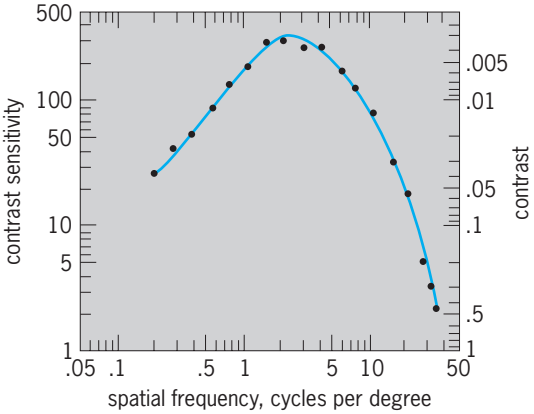


Fig. 7. Contrast sensitivity curve for a human subject. (*After F. W. Campbell and L. Maffei, Contrast and spatial frequency, Sci. Amer., 231:106–113, 1974*)

are more or less separate "channels" in the visual system for responding to spatial frequency, just as there are separate channels for color, motion, and orientation. (3) neurophysiologists have already identified single cells in the visual systems of cats and monkeys that are tuned for spatial frequency. Relatively coarse tuning is found in the cat, particularly with cells at the level of the retina or the thalamus. But in the monkey, especially in the visual cortex, some cells are found that respond only within narrow limits to particular grating frequencies. (4) the Fourier analysis of visual function is highly practical and efficient. It yields a quantitative comparison between human and animal capabilities in pattern vision. Similarly, it can be used to describe the course and severity of visual defect in patients with eye disease. Optical and lighting engineers are now using it to evaluate the conditions most favorable for vision throughout the entire frequency domain. And finally, any given pattern or scene can now be described in terms of its Fourier spatial frequency components, just as music or speech has long been analyzed in terms of its component temporal frequencies. Lorrin A. Riggs; Charles E. Sternheim

Bibliography. H. Davson, *The Physiology of the Eye*, 4th ed., 1990; R. Held, H. W. Leibowitz, and H. L. Teuber (eds.), *Perception*, vol. 8 of *Handbook of Sensory Physiology*, 1978; D. H. Hubel, *Eye, Brain and Vision*, 1995; D. Jameson and L. M. Hurvich (eds.), *Visual Psychophysics*, pt. 4 of vol. 7, *Handbook of Sensory Physiology*, 1972; R. Jung (ed.), *Central Processing of Visual Information*, pt. 3 of vol. 7, *Handbook of Sensory Physiology*, 1973; R. Sekuler and R. Blake, *Perception*, McGraw-Hill, Boston, 2002; B. A. Wandell, *Foundations of Vision*, 1995.

Visual debugging

Visualization of computer program state and program execution to facilitate understanding and, if necessary, alteration of the program. Debuggers are universal tools for understanding what is going on when a program is executed. Using a debugger, one can execute the program in a specific environment, stop the program under specific conditions, and examine or alter the content of the program variables or pointers. Traditional command-line oriented debuggers allowed only a simple textual representation of the program variables (program state). In the Unix GNU debugger (GDB), for instance, one would enter "print *variable*" to see a textual representation of *variable* on the console:

```
(gdb) print*tree
      *tree = {value = 7, name
            = 0x8049e88"Ada
```

This textual representation did not change even when modern debuggers came with a graphical user interface. Although variable names became accessible by means of menus, the variable values were still presented as text, including structural information, such as pointers and references. Likewise, the program execution is available only as a series of isolated program stops. (Pointers are variables that contain the "addresses" of other variables.)

This article presents the most important techniques of visual debugging, each exemplified by a specific tool.

Visualizing data structures. The GNU Data Display Debugger (DDD) is a graphical front-end to a command-line debugger such as GDB, providing menus and other graphical interfaces that eventually translate into debugger commands. As a unique feature, DDD allows the visualization of data structures as graphs. The concept is simple: Double-clicking on a variable shows its value as an isolated graph node. By double-clicking on a pointer, the dereferenced value or the variable pointed is to shown as another graph node, with an edge relating pointer and dereferenced value. By subsequent double-clicking on pointers, the programmer can unfold the entire data structure (**Fig. 1**).

If a pointer points to a value that is already displayed (for example, in a circular list), no new node is created; instead, the edge is drawn to the existing value. Using this alias recognition, the programmer can quickly identify data structures that are referenced by multiple pointers.

In principle, DDD can render arbitrary data structures by means of nodes and edges. However, the programmer must choose what to unfold, as the screen size quickly limits the number of variables displayed. Nonetheless, DDD is one of the most popular debugging tools under Unix and Linux.

Animating algorithms. Lens is an experimental graphical debugging front-end that includes an algorithm animation engine. The basic idea is to associate the code with algorithm animation directives. For instance, to visualize a tree, one would draw a square for a tree node containing the value and draw edges to the left and right children. These animation directives are then automatically executed during program execution.

The drawing and animation directives of Lens allow arbitrary visual rendering of data structures, providing an effective abstraction from the details of the implementation. Also, automatic unfolding of large data structures becomes feasible, which is not possible in DDD. Finally, programmers can also specify how transitions between program states should be rendered, thus allowing animated presentations.

The disadvantage, of course, is that for each data structure to be examined (and for each abstraction level), a specific visualization must be prepared, which requires more work than the one-size-fits-all graph visualization in DDD. Thus, Lens is more a presentation than an investigation tool; its presentations, though, have been used successfully in several classrooms to teach algorithms.

Summarizing executions. Jinsight is a tool for visualizing the dynamic behavior of Java programs. In contrast to Lens and DDD, Jinsight attempts to summarize entire program runs in its visualizations: First,

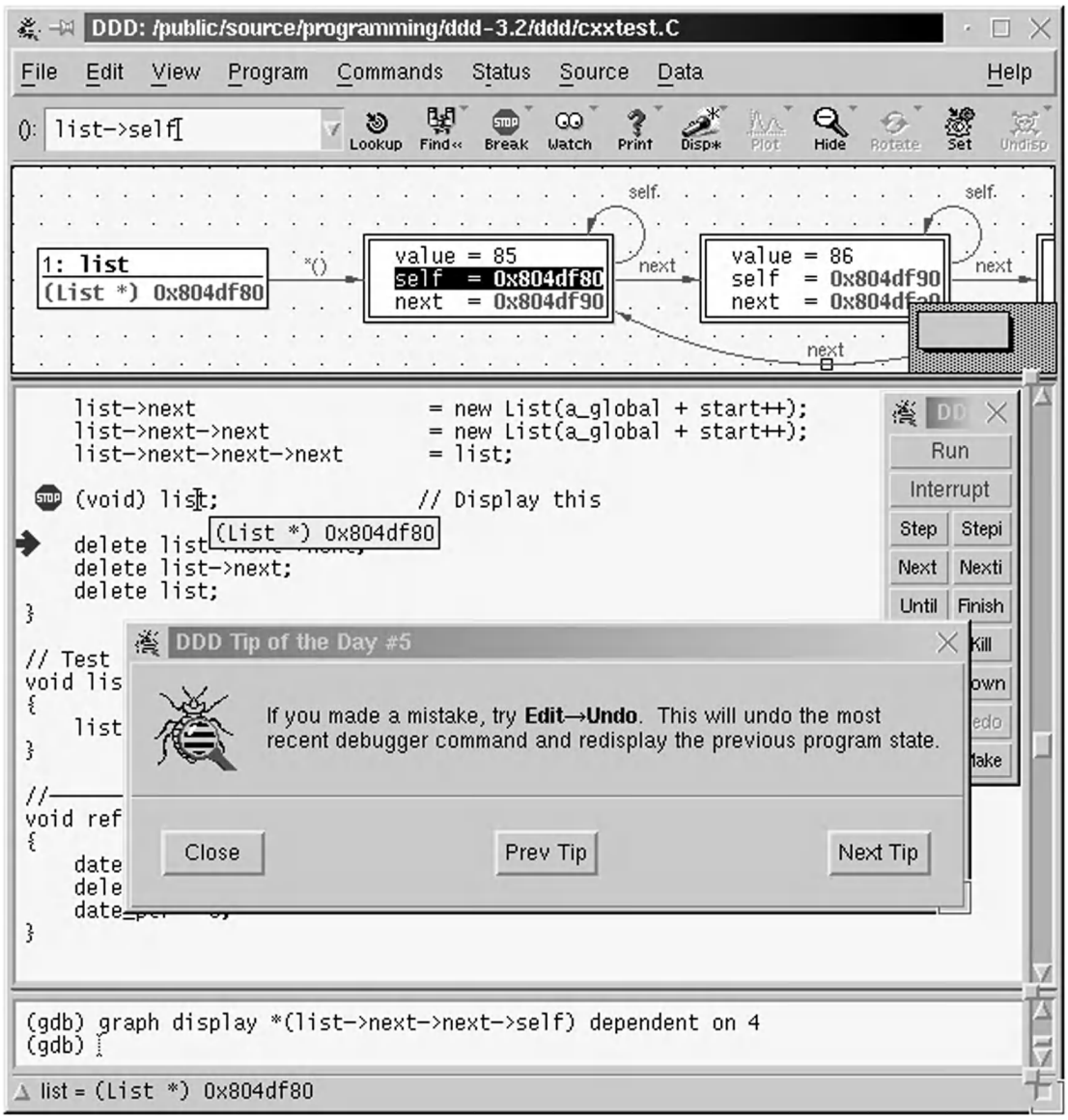


Fig. 1. Visualizing a list in DDD.

the program is executed using an extended Java virtual machine that tracks object events such as creation, interaction, or destruction during execution. At the end of the execution, this information is stored in a file which can then be examined by Jinsight.

Jinsight provides several views of the program execution. Object histograms show every single creation and destruction of arbitrary objects and highlight memory leaks. Call graphs show the sequence of invoked methods for several threads. (A thread is a portion of a program that can operate independently.) In all these views, Jinsight can identify common patterns and summarize these as well. For instance, in a call graph display, Jinsight can identify repeating sequences of function calls and summarize them to "this sequence of function calls is repeated n times." In practice, this effectively helps to isolate important events during program execution. Likewise, sets of objects with common properties can be summarized to a single visual representation, such as "each object of class c_1 references two objects of class c_2" (**Fig. 2**).

The drawback of Jinsight is that it allows only postmortem investigation; that is, the run can be neither explored nor influenced during execution. Also, Jinsight provides no means for exploring or rendering program states at one specific execution time. However, Jinsight is currently being extended to access program state while the program is still running, and will eventually be usable as an interactive debugger. Its innovative pattern recognition mechanisms make Jinsight the most advanced visual debugging tool available.

Conclusion. Compared to traditional debuggers, the techniques of visual debugging allow quicker exploration and understanding of what is going on in a program. Major challenges for the future include (1) finding appropriate techniques that summarize program executions and program states; (2) creating tools that can be used for both investigation and

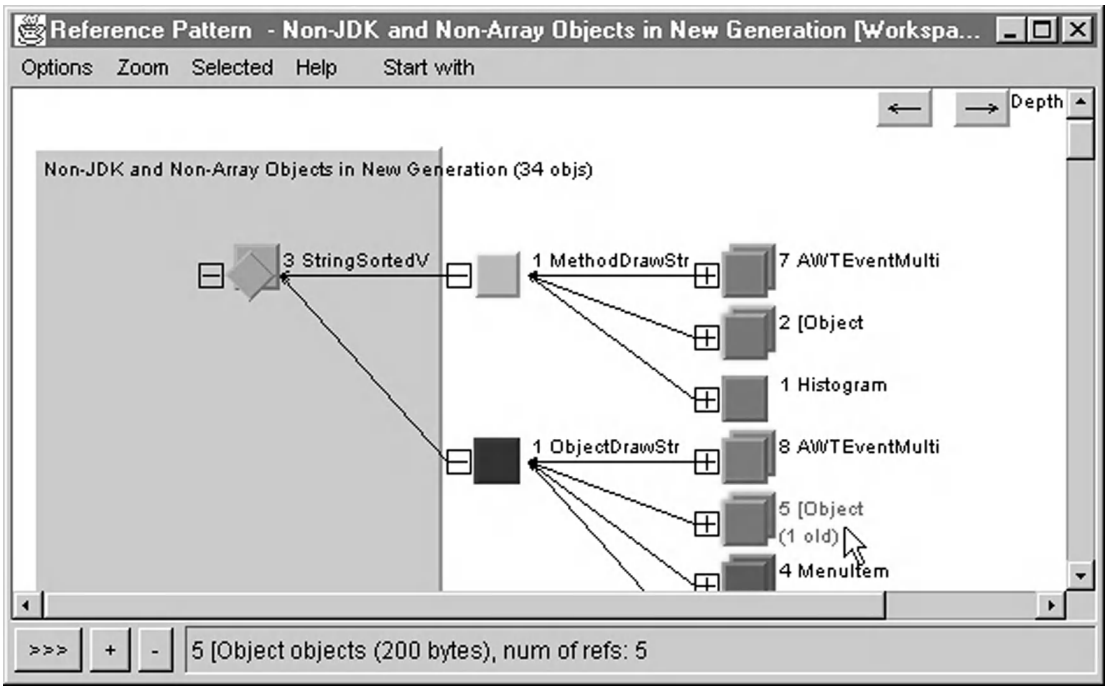


Fig. 2. Summarizing references in Jinsight.

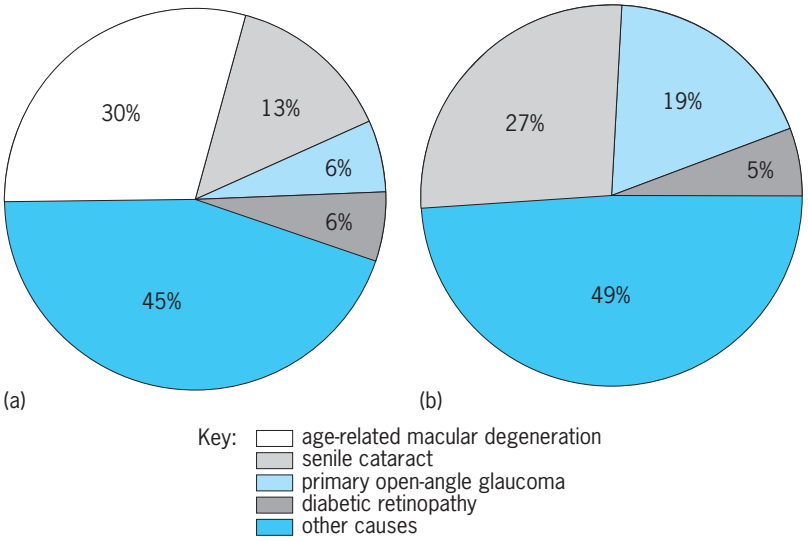
presentation; (3) creating tools that allow exploring specific states as well as sets of states. *See* COMPUTER PROGRAMMING; PROGRAMMING LANGUAGES; SOFTWARE; SOFTWARE ENGINEERING. Andreas Zeller

Bibliography. W. de Pauw et al., Visualizing the behavior of object-oriented systems, *OOPSLA'93 Conference Proceedings* (Washington, DC), pp. 326–337, September 1993; S. Mukherjea and J. T. Stasko, Toward visual debugging: Integrating algorithm animation capabilities within a source level debugger, *ACM Trans. Comput. Hum. Interac.*, 1(3):215–244, September 1994; A. Zeller, Visual debugging with DDD, *Dr. Dobbs J.*, no. 322, pp. 21–28, March 2001.

Visual impairment

Abnormal visual acuity. The term is used to describe visual acuity substantially less than normal. The World Health Organization defines visual impairment as acuity less than 20/60 (normal being 20/20). Most states in the United States, however, will not issue an unrestricted driver's license to persons having corrected vision less than 20/40; such persons generally require special aids to read fine print and to perform other visually demanding tasks. The legal definition of blindness in the United States is visual acuity of 20/200 or worse (or severely restricted peripheral vision). The World Health Organization defines blindness as visual acuity worse than 20/400.

Visual impairment and blindness increase substantially with age. Results of a survey of an urban United States population indicated that nearly 1% of the population is visually impaired at age 40–49; this percentage increases to 7–8% by age 80. The major causes of blindness differ substantially by race (see **illus.**). Cataracts, which involve opacification of the normally clear lens, and diabetic retinopathy, which is an accumulation of fluid or the growth of abnormal blood vessels in the retina (most commonly in insulin-dependent diabetics), are important causes of blindness. A large proportion of people having blindness caused by cataracts can be treated by surgery. A growing proportion of blindness caused by diabetic retinopathy can be prevented by laser surgery.



Major causes of blindness among Americans. (*a*) Whites. (*b*) Blacks. Glaucoma is more important among blacks; age-related macular degeneration is more important among whites. These preliminary data are from the Baltimore Eye Survey. Other causes of blindness include congenital cataract, corneal opacity, other types of glaucoma, hereditary retinopathy, macular scar, optic retinopathy, and trauma.

Other important causes of blindness are glaucoma and macular degeneration. Early detection of glaucoma requires routine, careful screening and examination, since many patients remain asymptomatic until much of the optic nerve is destroyed. Treatment may require drugs or surgery, the latter procedure creating an alternative pathway by which fluid can leave the eye, thereby lowering intraocular pressure to a safer level. Macular degeneration involves atrophy of that portion of the retina responsible for fine (reading) vision. A small proportion (probably less than 5%) of individuals with this condition also develop abnormal blood vessels that are amenable to laser surgery. For others, however, little can be done to prevent the onset or progression of age-related macular degeneration. However, while fine vision is often lost, peripheral vision is commonly retained. People with macular degeneration rarely suffer the total blindness of advanced glaucoma or diabetic retinopathy. *See* DIABETES; GLAUCOMA.

Visual problems are far more common in developing countries, where there are limited resources for dealing with problems that are otherwise readily treated, such as cataract, or prevented, such as trachoma or xerophthalmia ("blinding malnutrition" caused by vitamin A deficiency). In addition, in developing countries in the tropics, blinding parasitic diseases are common, particularly river blindness or onchocerciasis in Africa. Fortunately the outlook for onchocerciasis is improving with a successful, internationally organized campaign to control the black fly that spreads the disease and the development of a safer, more effective drug for treating the worm infestation that destroys vision. Trachoma, a chronic conjunctivitis that is endemic in many developing countries, is caused by the bacterium *Chlamydia trachomatis*. The infection can be treated with antibiotics or sulfonamides and prevented by proper hygiene and sanitation. *See* CATARACT; EYE DISORDERS; GLAUCOMA; VISION.

Alfred Sommer

Bibliography. H. Goldstein, *The Demography of Blindness Throughout the World*, Res. Ser. 26, American Foundation for the Blind, 1980; J. M. Last (ed.), *Maxcy-Rosenau Public Health and Preventive Medicine*, 13th ed., 1992; A. Sommer, Attack blindness!, *Amer. J. Ophthalmol.*, 102:387–389, 1986; J. M. Tielsch et al., Blindness and visual impairment in an American urban population: The Baltimore Eye Survey, *Arch. Ophthalmol.*, 108:286–290, 1990.

Vitamin

An organic compound required in very small amounts for the normal functioning of the body and obtained mainly from foods. Vitamins are present in food in minute quantities compared to the other utilizable components of the diet, namely, proteins, fats, carbohydrates, and minerals. The average human adult eats about 600 g (21 oz) of food per day on a dry-weight basis, of which less than a gram is vitamins.

The discovery of the vitamins has been the result of primarily two lines of investigation: the study of nutritional disease in people and the feeding of purified diets of known composition to experimental animals. In this way vitamin deficiency diseases, known as avitaminoses, have been described. With the advent of vitamin fortification of flour, cereals, and other foods, specific vitamin deficiency diseases such as scurvy, rickets, pellagra, and beriberi have become rare in most developed countries. However, marginal deficiencies of many of the vitamins, which are difficult to detect and largely ignored, are still relatively common, and there is increasing evidence that vitamin status is related to the risk of developing chronic disease.

Synthetic and natural vitamins usually have the same biological value. Different vitamins, which are often not related to each other chemically or functionally, are conventionally divided into a fat-soluble group (vitamins A, D, E, and K) and a water-soluble group [vitamin C (ascorbic acid) and the various B vitamins]. Some animals require inositol and choline, whereas for most species these are not vitamins. The vitamins, particularly the water-soluble ones, occur almost universally throughout the animal and plant kingdoms (see **table**). *See* CHOLINE; INOSITOL.

Variations in the dietary requirements for vitamins between different animal species are usually related to the varying synthesizing abilities of the animal's tissues or gastrointestinal flora; for example, some species can synthesize vitamin C, others cannot. In addition, vitamin absorption may be hampered by certain events within the intestinal tract. For example, uptake of the fat-soluble vitamins is generally impaired by diseases of the biliary tract, due in part to the lack of adequate emulsification that results from lack of bile salts. Thiaminases in some foods may destroy thiamin within the intestinal lumen. Biotin may be rendered unabsorbable in the presence of avidin, a protein (commonly found in raw egg whites) that binds biotin.

The B vitamins function as coenzymes that catalyze many of the anabolic and catabolic reactions of living organisms necessary for the production of energy; the synthesis of tissue components, hormones, and chemical regulators; and the detoxification and degradation of waste products and toxins. However, vitamin C and the fat-soluble vitamins do not function as coenzymes. Vitamins C and E and β-carotene (a precursor of vitamin A) act as antioxidants, helping prevent tissue injury from free-radical reactions. In addition, vitamin C functions as a cofactor in hydroxylation reactions. Vitamin D has hormonelike activity in calcium metabolism; vitamin A plays a critical role in night vision, growth, and maintaining normal differentiation of epithelial tissue; and vitamin K has a unique posttranscriptional role in formation of active blood-clotting factors. *See* ANTIOXIDANT; COENZYME.

Fat-Soluble Vitamins

The fat-soluble group of vitamins comprises vitamins A, D, E, and K.

Vitamin sources and deficiencies

Recommended name	Alternative name	Best sources	Deficiency	Daily RDA for adults*
Fat-soluble				
A, β-carotene and other carotenoids	Retinol and provitamin A (for carotene)	Fish liver oils; liver; dairy products; yellow, orange, and green plants; carrots; sweet potatoes	Poor growth and night vision; blindness	1 mg (5000 IU) (as retinol)
D	D_2 or ergocalciferol; D_3 or cholecalciferol	Fish oils, fortified dairy products	Rickets, osteomalacia	5 µg (200 IU)†
E	Tocopherol	Grains and vegetable oils	Neuropathy	10 mg (15 IU)
K	K_1 or menaquinone; K_2 or phylloquinone	Green vegetables	Bleeding	80 µg
Water-soluble				
Vitamin C	Ascorbic acid	Citrus fruits, fresh vegetables, potatoes	Scurvy	60 mg‡
Thiamine	B_1	Pork, liver, whole grains	Beriberi	1.5 mg
Riboflavin	B_2	Milk, egg white, liver, leafy vegetables	Cheilosis, glossitis	1.7 mg
Niacin	Nicotinic acid and nicotinamide (or niacinamide)	Yeast, wheat germ, meats	Pellagra	19 mg§
B_6	Pyridoxine	Whole grains, yeast, egg yolk, liver	Skin disorders; convulsion in infants	2.0 mg
Biotin	—	Liver, kidney, yeast	Skin disorders	30–100 µg
Folic acid	Folate	Liver, deep-green leafy vegetables	Macrocytic anemias	0.2 mg
B_{12}	Cobalamin	Liver, meats	Pernicious anemia	2.0 µg
Pantothenic acid	—	Liver, kidney, green vegetables, egg yolk		4–7 mg

*Recommended dietary allowance (RDA; 1989) for 25–50-year-old male.
†No requirement with adequate exposure to sunlight.
‡100 mg for smokers.
§60 mg of tryptophan can replace 1 mg of niacin.

Vitamin A. This vitamin occurs in two principal forms in nature: retinol, which is found only in animal sources, and certain carotenoids (provitamins), which are found only in plant sources. Carotenoids are the compounds that give many fruits and vegetables their yellow and orange colorings. The most abundant and best known of the carotenoids is *β*-carotene, a precursor of vitamin A. It is a provitamin, because its vitamin A activity occurs upon conversion to retinol within the body. Most of this conversion takes place in the intestine. Foods rich in *β*-carotene include carrots, yellow vegetables, and dark-green leafy vegetables (for example, spinach and broccoli), pumpkins, apricots, and melons. Preformed vitamin A or retinol is found in liver, egg yolk, fish, whole milk, butter, and cheese. Both retinol and *β*-carotene are soluble in fat but not in water, and are destroyed readily by oxidation, which causes significant losses during storage and cooking. *See* CAROTENOID.

Requirements for vitamin A are expressed as retinol equivalents (RE) in order to standardize the many different sources and types of vitamin A. Many nutrition labels still express vitamin A in International Units (IU), an earlier method of measuring vitamin A. The equivalent of 1000 micrograms of retinol (1000 RE) is 5000 IU, with a typical diet containing one-half of the vitamin A preformed and one-half from carotenoids.

Vitamin A is essential for vision, for adequate growth, and for tissue differentiation. One of the primary signs of vitamin A deficiency in animals is loss of appetite, accompanied by growth retardation. The process whereby cells and tissues become "programmed" to carry out their special functions is known as differentiation. Vitamin A is necessary for normal differentiation of epithelial cells. In vitamin A deficiency, cells lose their ability to differentiate properly, and because abnormal differentiation is part of the cancer process, scientists have been investigating whether vitamin A status might be involved in some way in the prevention of cancer. *See* CANCER (MEDICINE); VISION.

One of the earliest symptoms of vitamin A deficiency is night blindness, or an impaired ability to see in dim light. Severe deficiency produces partial or total blindness (xerophthalmia). Vitamin A deficiency is by far most widespread and most serious among young children, especially in developing countries. It is the leading cause of child blindness and, in combination with other factors such as protein-calorie malnutrition or infection, is associated with high rates of child mortality. Therapeutic doses of vitamin A are distributed in targeted areas of the world to prevent xerophthalmia and to treat people in whom the early stages of blindness have already occurred. Administration of vitamin A in developing countries has resulted in a remarkable reduction in mortality of preschool-age children.

Vitamin A as retinol is well tolerated in human

adults up to levels of 50,000 IU per day; however, this is far in excess of generally recognized needs. Pregnant women should limit their intake to 10,000 IU per day, since in animals excess vitamin A is known to cause birth defects. β-Carotene, on the other hand, is the safest form of vitamin A, because it is converted by the body in a regulated way. It is poorly absorbed from the gastrointestinal tract, and its conversion to retinol becomes progressively less efficient as vitamin A status improves. Thus, high intakes of β-carotene cannot lead to hypervitaminosis A; they can, however, produce a reversible orange-yellow coloration of the skin (hypercarotenemia), which is harmless. *See* VITAMIN A.

Vitamin D. Also known as calciferol, vitamin D is essential for the proper formation of the skeleton and for mineral homeostasis. Exposure of the skin to ultraviolet light catalyzes the synthesis of vitamin D_3 (cholecalciferol) from 7-dehydrocholesterol. The other major form of the vitamin, D_2 (ergocalciferol), is the product of the ultraviolet light–induced conversion of ergosterol in plants. Vitamin D is converted to 1,25-dihydroxyvitamin D in the liver and kidney. This biochemically active form of the vitamin promotes absorption of calcium from the gut and helps maintain blood calcium levels by increasing release of calcium from the bone. *See* CALCIUM METABOLISM.

In the United States, the major source of dietary vitamin D is fortified foods such as milk. The vitamin is stable in foods; storage and cooking do not appear to affect its activity. Even moderate amounts of exposure of the skin to sunlight can provide adequate synthesis of the vitamin, although dark-skinned persons and the elderly have more limited capacity to synthesize the vitamin.

In children, vitamin deficiency results in deformation of the skeleton (rickets); in adults, it leads to a weakening of the bone, resulting from demineralization (osteomalacia). These diseases are uncommon because of food fortification with vitamin D. However, rickets does occur in infants who are breast-fed without supplemental vitamin D or exposure to sunlight, and rickets and osteomalacia can both occur in patients with kidney failure since these people cannot convert vitamin D to 1,25-dihydroxyvitamin D. The elderly are also at higher risk of vitamin deficiency as a result of a more limited ability to synthesize vitamin D in the skin when exposed to the sun, and a less efficient conversion to the active form, than the young. Some evidence suggests that poor vitamin D intake is associated with an increased risk of osteoporosis (weakening of the bones in the elderly). *See* OSTEOPOROSIS.

Vitamin D is reasonably well tolerated in adults. Intakes below 50,000 IU/day are generally not toxic; however, vitamin D is potentially toxic, especially in young children, in whom intake should not exceed 2000 IU/day. There is generally little reason, however, to exceed 400 IU/day. *See* VITAMIN D.

Vitamin E. There are at least eight naturally occurring forms of vitamin E, known as tocopherols and tocotrienols. α-Tocopherol is the most common form and also has the highest biological activity. The richest sources of vitamin E are seed oils and the margarines and shortenings made from these oils. Nuts, seeds, whole grains, leafy green vegetables, eggs, and milk are also good sources.

Vitamin E functions primarily as an antioxidant protecting the body tissues from damaging reactions (peroxidation) that arise from many normal metabolic processes and exogenous toxic agents. Thus it protects biological membranes such as those found in the nerves, muscles, and cardiovascular system; helps to prolong the life of red blood cells; and helps the body use vitamin A optimally.

Dietary deficiency is rare in humans. However, a progressive neuromuscular disease is observed in children and adults who are severely vitamin E–deficient as a result of a variety of syndromes involving malabsorption of fat. Symptoms include loss of coordination and balance and, in severe cases, loss of the ability to walk. In adults, less acute deficiencies have resulted in shortened life for red blood cells and an accumulation of so-called age pigment in certain tissues. In premature infants, a commonly occurring vitamin E deficiency is associated with a hemolytic anemia, bleeding in the brain, and damage to the retina of the eye.

The requirement for vitamin E is related to the amount of polyunsaturated fat consumed in the diet. The higher the amount of this type of fat, the more vitamin E is required; however, sources of polyunsaturated fat are usually good sources of vitamin E (for example, seed and vegetable oils). *See* FAT AND OIL (FOOD).

Vitamin E has the ability to block the formation of nitrosamines, which are chemicals that can cause cancer. Whether vitamin E helps prevent certain cancers is still under study. Epidemiological studies do suggest that the risk of cataracts is reduced with higher intake of vitamin E, and an immune-enhancing effect of vitamin E has been observed in the elderly. *See* VITAMIN E.

Vitamin K. The two naturally occurring forms of this vitamin are K_1 (menaquinone), which is synthesized by bacteria, and K_2 (phylloquinone), which is synthesized by plants. They are fat-soluble and unstable to light. Vitamin K_3 (menadione) is a synthetic form. About half of the requirements for vitamin K comes from dietary sources, predominantly green leafy vegetables, and half is synthesized in the intestinal tract by bacteria.

Vitamin K is essential for the formation of prothrombin and at least five other proteins involved in the regulation of blood clotting. The vitamin plays an essential role in the carboxylation of glutamic acid to γ-carboxyl glutamyl residues in prothrombin and other clotting proteins. Vitamin K plays a similar role in the biosynthesis of other proteins found in the plasma, bone, and kidney. However, defective coagulation, and the resultant bleeding, is the only major sign of vitamin K deficiency.

Vitamin K deficiency has been reported in some newborn infants, because the intestinal bacteria do not begin to supply large enough amounts of the

vitamin until about one week after birth. For this reason, vitamin K is given immediately after birth to prevent bleeding. In addition, infant formulas are fortified with vitamin K. Deficiency of vitamin K in adults is extremely rare. It is usually caused by prolonged treatment with antibiotics that inhibit the growth of intestinal flora, or due to anticoagulant therapy. Vitamin K antagonists are used as anticoagulants and as rodenticides. *See* BLOOD; VITAMIN K.

Water-Soluble Vitamins

The water-soluble vitamins are vitamin C and the B vitamins, which include thiamine (vitamin B_1), riboflavin (vitamin B_2), vitamin B_6, niacin, folic acid, vitamin B_{12}, biotin, and pantothenic acid.

Vitamin C. This vitamin, also known as ascorbic acid, is a white crystalline compound that is highly soluble in water and a strong reducing agent. The stability of ascorbic acid decreases with increases in temperature and pH. Destruction by oxidation is a serious problem in that a considerable quantity of the vitamin C content of foods is lost during processing, storage, and preparation. Vitamin C is found mainly in citrus fruits and green vegetables (broccoli, brussels sprouts, spinach, kale, cabbage, cauliflower), but there are other sources such as cantaloupes, mangoes, strawberries, tomatoes, sweet peppers, and potatoes.

Of all the animals studied, only the guinea pig, the red-vented bulbul bird, certain fruit-eating bats, most fish, and the primates, including humans, require a dietary source of vitamin C. The other species studied are capable of synthesizing the vitamin in the liver or kidneys.

Vitamin C is required for the hydroxylation of the amino acids proline and lysine. This is a critical step for the production of collagen, the intercellular "cement" substance that gives structure to muscles, tissues, bones, and cartilage. The vitamin also plays an important role in the synthesis of a number of neurotransmitters, which are chemicals involved in nerve conduction, and in the synthesis of carnitine, a compound involved in the oxidation of fatty acids for energy. In addition, vitamin C enhances the absorption of iron from the diet, blocks the formation of potentially carcinogenic nitrosamines in foods and in the stomach, increases the resistance to infectious diseases, and is necessary for collagen formation and optimum wound healing. Finally, the vitamin can function as an antioxidant, preventing peroxidation in foods and in the body. *See* COLLAGEN.

Deprivation of vitamin C for a sufficient period of time results in scurvy, a potentially fatal disease characterized by weakening of collagenous structures, which results in widespread hemorrhaging of the capillaries. Bleeding gums and loosening of the teeth are usually the earliest objective signs. Hemorrhages under the skin cause extreme tenderness of extremities and pain during movement. If vitamin C is not administered, gangrene and death may ensue.

People who smoke have a higher requirement for vitamin C. Epidemiological studies indicate that consumption of foods high in vitamin C is associated with a lower risk of gastrointestinal cancers, higher levels of high-density-lipoprotein (HDL) cholesterol (cholesterol associated with decreased heart disease), and lower blood pressure. Whether vitamin C itself is responsible for these beneficial effects remains to be proven. Vitamin C has also been shown to reduce the duration and severity of symptoms of the common cold. The vitamin is used in foods and for industrial purposes as a reducing agent and antioxidant. *See* ASCORBIC ACID.

Thiamine. Thiamine, also known as vitamin B_1, is a water-soluble vitamin found in many foods; pork, liver, and whole grains are particularly rich sources. It is unstable to heat, alkali, oxygen, and radiation. Because thiamine is soluble in water, about 25% of the thiamine in food is lost during the normal cooking process. A number of foods such as coffee, tea, raw fish, betel nuts, and some cereals contain antagonists to vitamin B_1. Thiamine is also lost during certain food manufacturing processes, such as the polishing of rice and the refining of flour; therefore, these foods are generally enriched with thiamine.

Thiamine is essential for carbohydrate metabolism. In the form of thiamine pyrophosphate, it is a coenzyme required for the oxidative decarboxylation of α-keto acids and for the activity of transketolase in the pentose phosphate pathway, both key steps in the breakdown of glucose to energy. Thiamine also plays a role in the conduction of nerve impulses and in aerobic metabolism.

The two principal thiamine deficiency diseases are beriberi and the Wernicke-Korsakoff syndrome. Marginal deficiencies of thiamine may produce vague symptoms such as anxiety, hysteria, nausea, depression, and loss of appetite prior to the onset of clinical symptoms of beriberi. Beriberi was once endemic in areas where polished rice made up a large part of the diet, especially Southeast Asia. Many countries now fortify rice and other cereal grains to replace the nutrients lost in processing, which has resulted in a decrease in the incidence of beriberi. Today it is the Wernicke-Korsakoff syndrome, generally associated with alcoholism, that is more likely to be encountered. The deficiency is caused by a combination of factors, including inadequate intake of the vitamin (as when alcohol replaces food), decreased absorption, and increased thiamine requirements. Symptoms range from mild confusion and depression to psychosis and coma. If treatment is delayed, memory may be impaired permanently.

Because thiamine has a high turnover rate and is not appreciably stored in the body (approximately 1 mg/day is used up in tissues), a continual supply is required. The requirement for thiamine is related to energy intake, and increases as large amounts of carbohydrates are consumed. Similarly, periods of increased metabolism such as fever raise the thiamine requirement. *See* CARBOHYDRATE METABOLISM; THIAMINE.

Riboflavin. Also known as vitamin B_2, riboflavin is a water-soluble yellow-orange fluorescent pigment that is widely distributed in nature, occurring mostly in milk, egg white, liver, and leafy vegetables. It is

stable to acid and oxidation, but is rapidly destroyed by alkali at elevated temperatures and by light.

Riboflavin deficiency results in poor growth and other pathologic changes in the skin, eyes, liver, and nerves. Riboflavin deficiency in humans is usually associated with a type of cracking at the corners of the mouth known as cheilosis; inflammation of the tongue, which appears red and glistening (glossitis); corneal vascularization accompanied by itching; and a scaly, greasy dermatitis about the corners of the nose, eyes, and ears. Because riboflavin is essential for the functioning of vitamin B_6 and niacin, and B vitamins are generally found together in the same foods, it is often difficult to isolate riboflavin deficiency from those of other B vitamin deficiencies.

The vitamin functions biochemically in two coenzymes: flavin mononucleotide (FMN) and flavin adenine dinucleotide (FAD). Enzymes containing these coenzymes are known as flavoproteins. In general, these enzymes participate in oxidation-reduction reactions by accepting hydrogen ions (protons) and electrons from one substrate and transferring them to another.

Riboflavin requirements appear to be related to caloric requirements and muscular activity. *See* RIBOFLAVIN.

Vitamin B_6. This vitamin exists as three chemically related and water-soluble forms found in food: pyridoxine (pyridoxol), pyridoxal, and pyridoxamine. All three forms have equal activity for animals. They are all stable in acid but are rapidly destroyed by light in alkaline or neutral solutions. The best sources of vitamin B_6 are fish, chicken, organ meats, eggs, whole grains, beans, and nuts. Considerable losses occur in freezing of fruits and vegetables, processing of luncheon meats, and milling of cereals.

The three forms of vitamin B_6 are converted in the liver, red blood cells, and other tissues to pyridoxal phosphate and pyridoxine phosphate, which serve as coenzymes in transamination and decarboxylation reactions and other metabolic transformations of amino acids and in the metabolism of fats and nucleic acids. The vitamin is essential for the conversion of tryptophan to niacin.

Vitamin B_6 deficiency rarely occurs alone; it is commonly seen in people who are deficient in several B-complex vitamins. Clinical signs of deficiency include epileptiform convulsions, dermatitis, and anemia. Deficiency in adults is associated with depressed immune response and mental depression; in infants it leads to a variety of neurological symptoms as well as abdominal distress.

Approximately 40 drugs (for example, oral contraceptives, isonicotinic acid hydrazid, and penicillamine) are known to interfere with its metabolism. The need for pyridoxine increases proportionately with protein intake. Inadequate intake of vitamin B_6 is common, particularly in women, the elderly, alcoholics, and diabetics. Prolonged use of 500 mg or more per day of vitamin B_6 can cause ataxia and a severe sensory neuropathy. *See* VITAMIN B_6.

Niacin. Niacin, also known as nicotinic acid, is a white water-soluble powder stable to heat, acid, and alkali. It is found in biochemically active combinations as the amide, niacinamide. All living cells that have been studied have enzymic systems involving niacin. Many animals, including humans, are capable of synthesizing niacin in varying degrees from the amino acid tryptophan. Niacin is widely distributed in foods: yeasts, wheat germ, and meats, particularly organ meats, are rich sources of the vitamin. Some foods such as milk are relatively poor sources of niacin, but contain generous quantities of tryptophan.

Niacin-deficiency disease is known as pellagra and is particularly prevalent among people whose diet is largely corn. Pellagra is actually a multiple-deficiency disease; some cases respond to niacin alone while others respond only to mixtures of niacin and other B vitamins.

The pellagra-preventive potency of a diet is related not only to its niacin and tryptophan content but also to the actual availability of its niacin. There is evidence that some of the niacin content of foods cannot be released by digestive enzymes. The existence of an antiniacin material in corn has been suggested. Very high levels of niacin are used to lower blood cholesterol, but this type of treatment must be monitored carefully for side effects, such as flushing and liver damage. *See* NIACIN.

Folic acid. Folate and folacin are terms for compounds that have nutritional properties and chemical structures similar to those of folic acid (pteroylglutamic acid; PGA). Folates function metabolically as coenzymes that transport single carbon fragments from one compound to another in amino acid metabolism and nucleic acid synthesis. Deficiency of the vitamin leads to impaired cell division and to alterations of protein synthesis—effects most noticeable in rapidly growing tissues.

Folate is widely distributed in foods; liver, yeast, leafy vegetables, legumes, and some fruits are especially rich sources. As much as 50% of food folate may be destroyed during household preparation, food processing, and storage. Most dietary folates exist in the polyglutamate form (that is, with glutamic acids attached), which is converted in the wall of the small intestine to the monoglutamate form before it is absorbed into the bloodstream. The bioavailability of folate in a typical diet in the United States is about one-half that of crystalline folic acid, which is efficiently absorbed.

Deficiency is frequently characterized by megaloblastic anemia, a disease marked by the presence of abnormal red blood cells in the bone marrow, and macrocytic anemia, a disease marked by oversized red blood cells. Vitamin B_{12} deficiency causes a similar type of anemia that is responsive to folic acid. Therefore, administration of folate could mask a deficiency of vitamin B_{12}, which could be manifested later as neurological damage. Poor folic acid status is associated with increased birth defects (neural tube defects and cleft palate) in the fetus and the newborn. Use of a multivitamin supplement containing folic acid just before and during the first weeks of pregnancy has been shown to reduce the risk of

such defects substantially. Abnormal tissue development (dysplasia), characteristic of some precancerous conditions, can also be treated by folate supplementation. Poor folic acid status is prevalent in the newborn, adolescent females, pregnant women, and the elderly. *See* ANEMIA; FOLIC ACID.

Vitamin B_{12}. The terms vitamin B_{12} and cobalamin refer to all members of a group of large cobalt-containing corrinoids that can be converted to methylcobalamin or 5′-deoxyadenosylcobalamin, the two cobalamin coenzymes active in human metabolism. Cyanocobalamin is the commercially available form of vitamin B_{12} used in vitamin supplements and pharmaceuticals. This form is soluble in water and stable to heat. When given either orally or parenterally, it is converted by the removal of cyanide to the forms that are metabolically active in humans. In plasma and tissue, the predominant forms are methylcobalamin, adenosylcobalamin, and hydroxocobalamin.

Bacteria, fungi, and algae can synthesize vitamin B_{12}, but yeasts, higher plants, and animals cannot. In the human diet, vitamin B_{12} is supplied primarily by animal products, where it has accumulated from bacterial synthesis. Plant foods are essentially devoid of vitamin B_{12}.

Vitamin B_{12} deficiency results in macrocytic, megaloblastic anemia, in neurological symptoms due to demyelination of the nerves, and in other less specific symptoms (such as sore tongue or weakness). Neuropsychiatric manifestations of vitamin B_{12} deficiency are seen in the absence of anemia, particularly in the elderly. It is rare for a diet to be deficient in vitamin B_{12}; most of the vitamin B_{12} deficiency seen in the United States is due to inadequate absorption. Deficiency can be produced by strict vegetarian diets devoid of meat, eggs, and dairy products.

The intestinal absorption of vitamin B_{12} is mediated by a highly specific binding glycoprotein (Castle's intrinsic factor), which is secreted in the stomach. In pernicious anemia, vitamin B_{12} is not absorbed from the diet or reabsorbed from the bile due to lack of intrinsic factor activity. Vitamin B_{12} is generally injected into individuals with pernicious anemia or those who have undergone gastrectomy to prevent anemia and neuropathy. *See* VITAMIN B_{12}.

Biotin. This is a sulfur-containing B-complex vitamin that is very widespread in nature. Biotin is only sparingly soluble in water; it is stable in boiling water solutions. It can be destroyed by oxidizing agents, acids, and alkalies, and under some conditions by oxidation in the presence of rancid fats. Biotin functions as a coenzyme in carbon dioxide fixation and decarboxylation enzymes; it plays an important role in metabolism of carbohydrates, fats, and proteins.

Raw egg white contains a protein (avidin) that binds biotin, rendering it unavailable. Thus biotin deficiency can be induced by the ingestion of large amounts of avidin. This condition is characterized by anorexia, nausea, vomiting, mental depression, hair loss accompanied by a dry scaly dermatitis, and an increase in serum cholesterol and bile pigments.

Dietary deficiency of biotin is very rarely encountered except in infants. A nutritional deficiency of biotin that has been observed in infants under 6 months of age is characterized by skin disorders (seborrheic dermatitis). *See* BIOTIN.

Pantothenic acid. A member of the B-vitamin group, pantothenic acid is a light-yellow, viscous oil that is readily soluble in water; it is usually obtained as the calcium salt. Pantothenic acid is widely distributed; among its best sources are liver, kidneys, fresh green vegetables, and egg yolks. Loss of the vitamin during cooking is minimal, as it is present in stable conjugated form in food.

Pantothenic acid plays its primary physiological roles as a component of the coenzyme A molecule and of the acyl carrier protein of fatty acid synthetase. Reactions with these substances are important in the release of energy from carbohydrates; in gluconeogenesis; in the synthesis and degradation of fatty acids; and in the synthesis of such vital compounds as sterols and steroid hormones, porphyrins, and acetylcholine.

Evidence of dietary deficiency in humans has not been recognized clinically; however, experimentally produced deficiency has resulted in headache, fatigue, impaired motor coordination, paresthesia (prickling or tingling of the skin), muscle cramps, and gastrointestinal disturbances. A syndrome known as burning feet, which was observed among prisoners of war and malnourished subjects in the Far East, is believed by some authorities to represent a specific deficiency in humans, since it responds to pantothenic acid but not the other B vitamins. Pantothenic acid deficiency is difficult to identify in humans, because it is usually accompanied by general malnutrition and deficiency of the other B vitamins.

A recommended dietary allowance (RDA) has not been set, but an intake of 4–7 mg for an adult is considered safe and adequate. *See* NUTRITION; PANTOTHENIC ACID.

Lawrence J. Machlin

Bibliography. A. Bendick and C. E. Butterworth (eds.), *Micronutrients in Health and in Disease Prevention*, 1991; W. Friedrich, *Vitamins*, 1988; L. J. Machlin, *Handbook of Vitamins*, 2d ed., 1990; National Research Council, *Recommended Dietary Allowances* 10th ed., 1989; M. E. Shils and V. R. Young, *Modern Nutrition in Health and Disease*, 9th ed., 1999.

Vitamin A

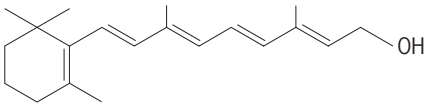
A pale-yellow alcohol, soluble in fat but not in water. In pure form, it is readily destroyed by oxidation and light, which may cause losses during storage.

Source. Vitamin A is found in all animal tissues, although it is particularly concentrated in the liver. There are two different dietary sources for the vitamin: animal sources which contain vitamin A itself, mostly in the form of retinyl esters, and plant sources which contain carotenoids that are converted to

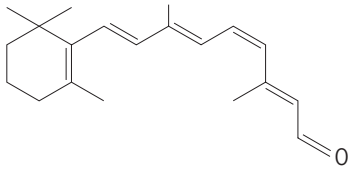
vitamin A in animal tissues such as the absorptive cells in the intestine. The most vitamin A–enriched animal food source is fish liver oil. Plant carotenoids are found in green and yellow fruits and vegetables such as carrots, apricots, asparagus, broccoli, and green leafy vegetables. Some fresh-water fish contain vitamin A_2, which differs slightly from vitamin A in structure. *See* CAROTENOID.

Bioassay. The vitamin A activity of carotenoids varies with their chemical structure, with β-carotene being the most potent. One international unit (IU) of vitamin A has been set at 0.3 microgram of vitamin A or 0.6 μg of β-carotene. This is somewhat confusing since the efficiency of conversion of β-carotene to vitamin A becomes greater in a deficiency state. Although biological assays are sometimes used, vitamin A and carotene are usually determined by spectrophotometric techniques. *See* MASS SPECTROMETRY.

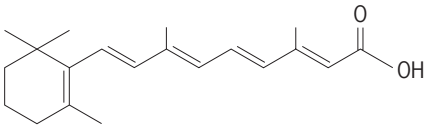
Physiological activities. In vitamin A deficiency, the epithelial tissues of many organs are affected. Growth failure occurs, and young animals can suffer from neurological symptoms resulting from pressures on the central nervous system. Changes occur in the skin, mouth, respiratory tract, urogenital tract, and some glands. Vitamin A deficiency is also strongly associated with depressed immune function and higher morbidity and mortality due to infectious diseases such as diarrhea, measles, and respiratory infections. A severe manifestation of vitamin A deficiency is night blindness and inflammation of the eyes (xerophthalmia), followed by irreversible blindness. The symptoms seen in vitamin A deficiency reflect the multiple roles of this compound in animals. These roles are fulfilled by two compounds that are synthesized from vitamin A in the body: vitamin A aldehyde (retinaldehyde), which is critical for vision, and vitamin A acid (retinoic acid), which controls many physiological functions in both the embryo and the adult. The similar chemical structures of vitamin A, retinaldehyde, and retinoic acid are shown at right.



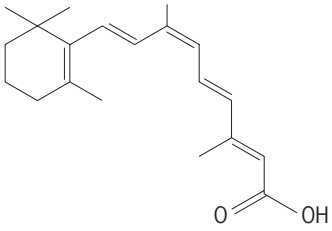
Vitamin A: all-*trans*-retinol



11-*cis*-Retinaldehyde



All-*trans*-retinoic acid



9-*cis*-Retinoic acid

Retinaldehyde. Retinaldehyde has a critical role in sight. This compound binds to a protein termed opsin to generate the visual pigment rhodopsin in the retina. Rhodopsin is the "visual antenna" that responds to light by sending signals to the brain. Following absorption of light, rhodopsin releases the bound vitamin A compound. Resynthesis of rhodopsin by providing fresh molecules of retinaldehyde is therefore essential for normal vision. *See* VISION.

Retinoic acid. Retinoic acid is the active vitamin A metabolite that is critical for many processes, such as cell growth and differentiation, in the embryo and in the adult. Retinoic acid exerts its effects by modulating the expression of many genes. This activity is due to the ability of the compound to control the activities of proteins known as retinoid nuclear receptors. When activated by retinoic acid, these proteins bind to deoxyribonucleic acid (DNA) and regulate gene transcription. Retinoic acid is currently used in treatment of various pathologies such as skin disorders and certain types of cancer. *See* DEOXYRIBONUCLEIC ACID (DNA).

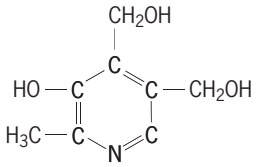
Requirements. Studies of many mammalian species suggested that approximately 20 IU (6 μg) of vitamin A per kilogram of body weight will support growth and prevent symptoms of deficiency. The current intake recommendations of vitamin A in the United States is 3 mg/day and about 1 mg/day in the European Union. Vitamin A deficiency is a serious public health problem in many countries but is essentially nonexistent in western Europe and the United States. The excessive intake of vitamins has resulted in cases of vitamin A toxicity in people receiving 25 to several hundred times the requirement over prolonged periods. Discontinuance of vitamin A supplementation in these individuals is usually followed by a dramatic disappearance of symptoms of toxicity in a few days. *See* VITAMIN.

Stanley N. Gershoff; Noa Noy

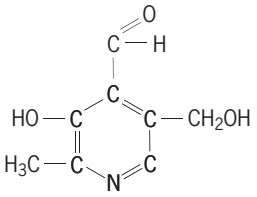
Bibliography. D. Mangelsdorf, K. Umesono, and R. M. Evans, The retinoid receptors, in M. B. Sporn et al. (eds.), *The Retinoids*: *Biology*, *Chemistry*, *and Medicine*, Raven Press, New York, 1994; N. Noy, Vitamin A and carotenoids, in M. H. Stipanuk (ed.), *Biochemical and Physiological Bases of Human Nutrition*, W. B. Saunders, Philadelphia, 1999; S. A. Ross et al., Retinoids in embryonal development, *Physiol. Rev.*, 80:1021–54, 2000; R. D. Semba., Vitamin A, immunity and infection, *Clin. Infect. Dis.*, 19:489–499, 1994.

Vitamin B_6

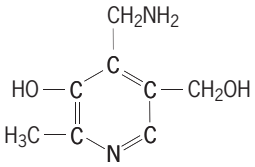
A vitamin which exists as three chemically related and water-soluble forms found in food, as shown here. All three forms have equal activity for animals



Pyridoxine (pyridoxol)



Pyridoxal



Pyridoxamine

and yeast, but pyridoxal and pyridoxamine have several thousand times the activity of pyridoxine for some bacteria. Vitamin B_6 is stable in acid, but is rapidly destroyed by light in alkaline or neutral solutions.

Bioassay. Chemical assays for vitamin B_6 are laborious; microbiological assays are commonly used. They often give results which appear much too high when compared to animal-feeding tests. This is because vitamin B_6 is conjugated with other substances in foods, and the acid hydrolysis procedures used in preparing samples for microbiological assay release quantities of the vitamin not released by normal digestive processes. *See* BIOASSAY.

Dietary requirements. Most studies of vitamin B_6 deficiency have been done on animals, since a deficiency of this vitamin in humans is very rare. Vitamin B_6 deficiency is accompanied by poor growth, dermatitis, microcytic anemia, epileptiform convulsions, and kidney and adrenal lesions. There is evidence that some women in the third trimester of pregnancy may have a special requirement for vitamin B_6 in that its administration often relieves the nausea of pregnancy. Some types of human dermatitis respond to local application of this vitamin. Experimental studies have suggested that vitamin B_6–deficient animals are deficient in available insulin and growth hormone. Marked abnormalities in lipid, carbohydrate, and protein metabolism are caused by this deficiency.

Vitamin B_6–deficient animals are oxaluric and may form kidney stones. Atherosclerosis and impaired antibody formation have been reported. Vitamin B_6–responsible anemias have been reported in humans, and it has been suggested that supplemental vitamin B_6 may protect against dental caries, particularly during pregnancy.

It is difficult to set requirements for vitamin B_6, since no single set of assay conditions or criteria has received universal acceptance. Based on animal experiments, a number of dietary factors probably affect the vitamin B_6 requirement. High-protein diets increase the requirement. High-carbohydrate diets probably increase intestinal synthesis of vitamin B_6, while unsaturated fatty acids decrease the requirement. Adults probably require about 1.5–2 mg per day, and a dietary intake of 0.4 mg per day would probably be satisfactory for most infants.

Biochemistry. Vitamin B_6 functions as a coenzyme in the form of pyridoxal 5-phosphate. The value of pyridoxine and pyridoxamine lies in the ability of tissues to convert them into pyridoxal. Vitamin B_6–containing enzymes are included in aminotransferase reactions (transamination), important mechanisms for the synthesis of amino acids by the tissues from α-keto acids. An example is shown below.

Glutamic acid + pyruvic acid $\rightleftharpoons$

α-ketoglutaric acid + alanine

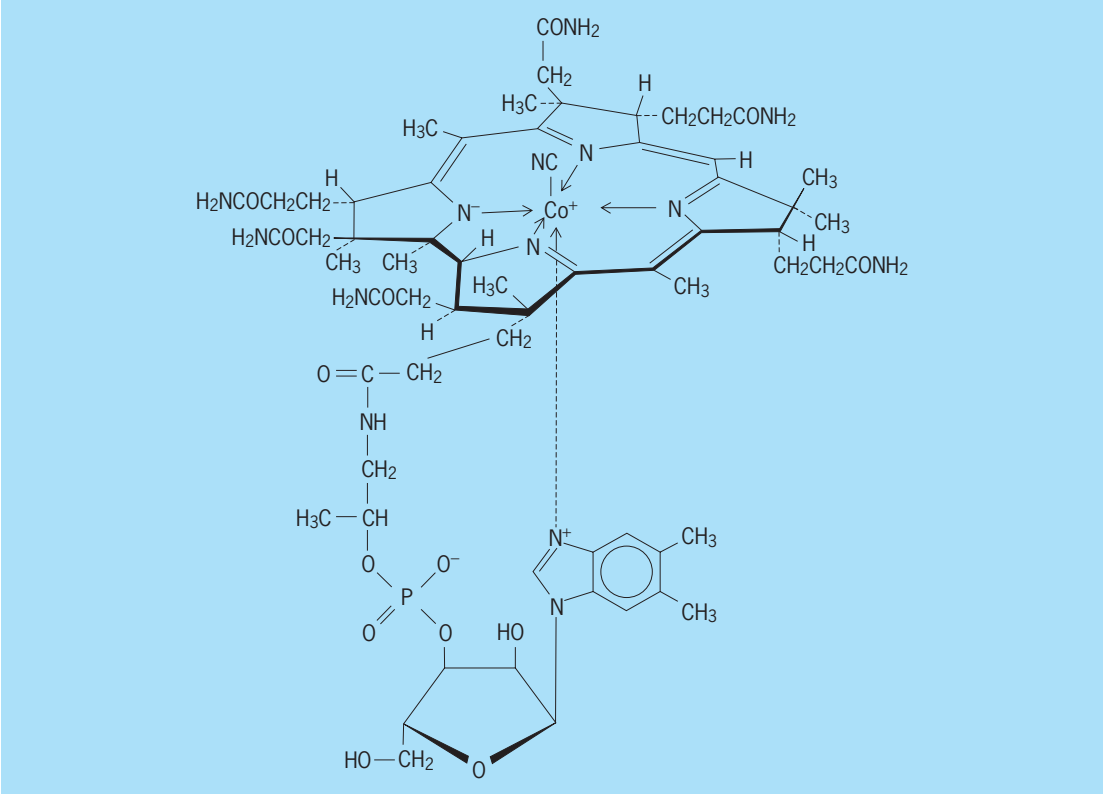
Vitamin B_6 is also involved in amino acid decarboxylation reactions. The transformations of histidine to histamine and of aspartic acid to β-alanine are examples of these. This vitamin has a special function in the metabolism of tryptophan and is necessary for the conversion of tryptophan to niacin. In most vitamin B_6–deficient animals, xanthurenic acid, a metabolite of tryptophan, is excreted in abnormal quantities, and this has been used as the basis for tests of the adequacy of vitamin B_6 nutrition. This vitamin also has a special function in the metabolism of the sulfur-containing amino acids. *See* COENZYME; TRANSAMINATION; VITAMIN. Stanley N. Gershoff

Bibliography. M. H. Briggs, *Vitamins in Human Biology and Medicine*, 1981; S. S. Hendler, *Vitamin and Mineral Encyclopedia,* 1990; G. B. Tryfiates, *Vitamin B_6: Metabolism and Its Role in Growth*, 1980.

Vitamin B_{12}

A group of closely related polypyrrole compounds containing trivalent cobalt (see **illus.**); often called cobalamin. The vitamin is a dark-red crystalline compound; in aqueous solution and at room temperature it is most stable at pH 4–7.

Source and assay. In general, vitamin B_{12} is synthesized by microorganisms, not by plants, and is found in animal tissues as a result of intestinal synthesis or ingestion. It is now thought that the animal-protein factor, which was studied widely, is at least in part vitamin B_{12}. Vitamin B_{12} is usually determined microbiologically, although rat and chick growth assays are also available. *See* BIOASSAY.



Structural formula for vitamin B_{12}.

Physiology and requirements. Vitamin B_{12} deficiency in animals is characterized primarily by anemia and neuropathy. In humans, this deficiency is called pernicious anemia. People suffering from this disease lack a factor secreted in normal gastric juice which, by affecting absorption directly and by protecting vitamin B_{12} from intestinal destruction, enables the vitamin to be absorbed. This factor, called intrinsic factor, appears to be one or more mucoproteins. Its action in facilitating the absorption of vitamin B_{12} is calcium-dependent. Vitamin B_{12} deficiency in humans may be caused by a genetically controlled lack of intrinsic factor, gastrectomy, malabsorption disease, or vegetarianism. Body stores of the vitamin are such that it takes approximately 5 years for a deficiency state to appear clinically following gastrectomy.

Requirements for vitamin B_{12} are increased by reproduction or hyperthyroidism. Of the known vitamins, B_{12} is the most active biologically. A daily injection of 1 microgram of vitamin B_{12} will prevent the recurrence of symptoms in people with pernicious anemia. For normal people a diet containing 3–5 μg per day (providing 1–1.5 μg absorbed) will satisfy vitamin B_{12} requirements. Some of the vitamin B_{12} requirement is probably met by intestinal synthesis.

Biochemistry. Several coenzymes of vitamin B_{12} exist. They are included in isomerization reactions (methylmalonyl-CoA mutase, glutamate mutase), conversion by dioldehydrases of 1,2-diols to deoxyaldehydes, methylation reactions (formation of methionine and thymidine 5′-phosphate), and conversion of ribose to deoxyribose. *See* VITAMIN.

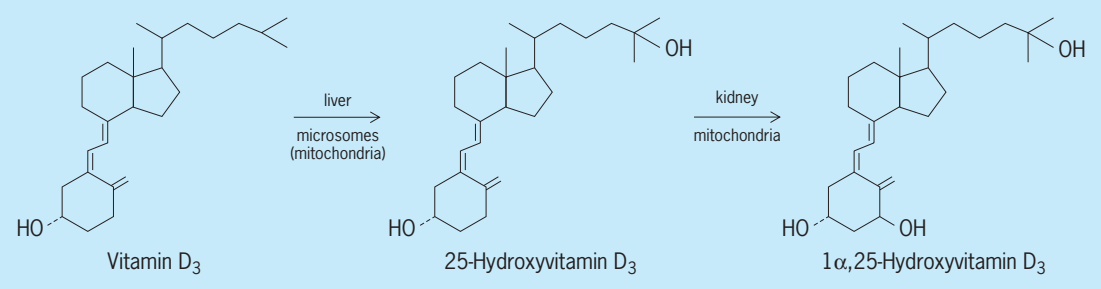
Stanley N. Gershoff

Bibliography. S. S. Hendler, *Vitamin and Mineral Encyclopedia*, 1990.

Vitamin D

Either of two fat-soluble sterol-like compounds, ergocalciferol (vitamin D_2) and activated cholecalciferol (vitamin D_3). Vitamin D_2 is formed from the irradiation of ergosterol, a plant sterol. However, vitamin D_3 is normally manufactured in the skin, where ultraviolet light activates the compound 7-dehydrocholesterol. Vitamins D_3 and D_2 are about equal in activity in all mammals except New World monkeys and birds, in which vitamin D_2 is approximately one-tenth as active as vitamin D_3. *See* VITAMIN.

Assay. In crystalline form, vitamin D compounds are determined by their ultraviolet absorption maximum by means of spectrophotometers. However, at concentrations normally found in foods and some vitamin pills, this method is not sufficiently sensitive. Generally, high-performance liquid chromatography is used to determine vitamin D levels in high-potency pills and other concentrates and food. Adequacy of vitamin D production in skin is assessed by the measurement of its major circulating metabolite, 25-hydroxyvitamin D. This is done by



Conversion of vitamin D to its active form in the body. Vitamin D_3 can be obtained from the diet, or it can be produced in skin by the ultraviolet irradiation of 7-dehydrocholesterol. 1α,25-Dihydroxyvitamin D_3 is believed to be the only active form of vitamin D in the body.

immunoassay, radioreceptor assay, or high-performance liquid chromatography. The active form of vitamin D, 1, 25-dihydroxyvitamin D_3, can be determined only by micromethods, such as radioreceptor assay, radioimmunoassay, or reporter gene cellular assay. Normal plasma levels of vitamin D_3 are 1–2 μg/mL of 25-hydroxyvitamin D, 15–20 μg/mL and of 1,25-diphydroxyvitamin D_3, 35–40 picograms/mL. *See* BIOASSAY; IMMUNOASSAY; LIQUID CHROMATOGRAPHY; RADIOIMMUNOASSAY.

Biochemistry. Vitamin D as acquired from the diet or produced in the skin is biologically inactive. It must be metabolized by the liver to produce 25-hydroxyvitamin D_3. However, this compound is also biologically inactive under physiological circumstances and must be activated by the kidney to produce the final vitamin D hormone, 1,25-dihydroxyvitamin D_3 (see **illus.**). This hormonal form of vitamin D plays an essential role in stimulating intestinal absorption of calcium and phosphorus, in the mobilization of calcium from bone, and in renal reabsorption of calcium. This hormone acts very much like other steroid hormones in that it binds to a nuclear receptor that stimulates the cell to produce proteins that carry out the functions of vitamin D. The molecular details of the mechanisms that these proteins utilize to increase absorption of calcium and phosphorus, mobilize calcium from bone, and activate renal reabsorption of calcium remain under investigation.

The function of vitamin D has been expanded beyond regulating plasma calcium and phosphorus levels, and hence healing the diseases of rickets and osteomalacia. It is now known that the vitamin D hormone controls parathyroid gland growth and production of the parathyroid hormone. It is an immunomodulator and is being developed for use in autoimmune diseases. Vitamin D hormone also appears to play a role in the regulation of insulin production or secretion. It is also used to treat the skin disorder psoriasis. These new sites of action of vitamin D are under intense investigation. *See* HORMONE.

Sources. Vitamin D is largely absent from the food supply. It is found in large amounts in fish liver oils; cod liver oil has long been known to be an important source of vitamin D. Except for unusual circumstances, it is not found in the plant world. It is in low concentrations in milk that is not fortified and in eggs because it is stored in the egg yolk to support growth and development of the chicken embryo. Fortified foods are the major dietary source of vitamin D, but the major overall source is the production of vitamin D in skin by exposure to sunlight or ultraviolet irradiation. In winter months at temperate latitudes, insufficient amounts of vitamin D are produced in skin, and unless it is replaced by a dietary source, danger of insufficiency exists.

Deficiency. A deficiency of vitamin D in growing animals results in the disease rickets, which is marked by a failure to mineralize the skeleton because of insufficient amounts of calcium and phosphorus in the plasma due to insufficient absorption from the gut and insufficient reabsorption in the kidney. A similar disease, osteomalacia, occurs in adult animals. Osteomalacia is characterized by a failure to calcify newly forming organic matrix of bone. By far the most serious disorder of vitamin D deficiency is the low-blood calcium levels which result in convulsions known as hypocalcemic tetany. Under circumstances of low-blood calcium levels, nerves continuously excite muscle, giving rise to continuous muscle contractions or convulsions. Moderate deficiency of vitamin D may contribute to osteoporosis, especially in the elderly. *See* BONE; OSTEOPOROSIS.

Toxicity. When doses of vitamin D exceed 5000–10,000 IU per day, danger of intoxication is present. Intoxication results from high plasma calcium concentrations that lead to calcification of kidney, heart, aorta, and subcutaneous tissues. Large doses of vitamin D can be lethal. Doses above 800 IU per day should never be taken unless prescribed by a physician.

Dietary requirements. The recommended daily requirement for vitamin D_3 is 10 micrograms or 400 international units (IU). Higher requirements are reported for the elderly and for rapidly growing adolescents: 20 μg or 800 IU per day. It is possible that the average requirement is lower than 10 μg per day. The exact absolute requirement has never been determined.

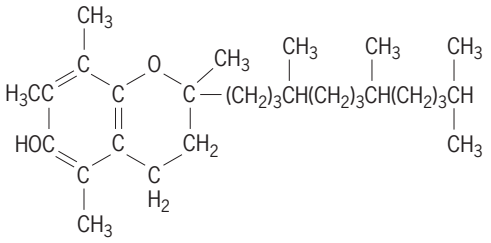
Therapeutic applications. The major applications of 1,25-dihydroxyvitamin D_3 and its analogs are in treatment of the bone disease associated with kidney

failure. It is also used in many countries to treat the bone-thinning disease, osteoporosis. For example, in Japan, it and its analog, 1α-hydroxyvitamin D_3, are the major treatments for osteoporosis. It is also used for the treatment of some genetic vitamin D–resistance diseases and for patients who lack parathyroid glands. More recent applications are for certain autoimmune diseases and for the treatment of the skin disorder psoriasis. Hector F. Deluca

Bibliography. D. Feldman, F. H. Glorieux, and J. W. Pike (eds.), *Vitamin D*, Academic Press, San Diego, 1997.

Vitamin E

A group of compounds, α, β, γ, and δ tocopherols, that have a chromanol ring and phytyl side chain and are widely distributed in nature, especially in edible vegetable oils (wheat germ, sunflower, cottonseed, safflower, canola, soybean, and corn oil). Unprocessed grains, nuts, and vegetables are other sources. When used as supplements, there are two vitamin E products available: natural source RRR α-tocopherol and synthetic all-rac-α-tocopherol. The



α-Tocopherol

latter is a mixture of eight different stereoisomers, of which only one is RRR α-tocopherol.

Physiological activities. The terms "vitamin E" and "α-tocopherol" are frequently used interchangeably in human nutrition, but it is imperative to distinguish between supplements of RRR α-tocopherol and of synthetic α-tocopherol because their biological activity is different. Since the major function of vitamin E is to serve as a chain-breaking antioxidant, protecting cell membranes against free-radical damage, the most potent form of the vitamin should be used as a supplement. Although gastrointestinal absorption of all forms of vitamin E is equivalent, the subsequent physiological steps are sharply in favor of the RRR form. This action is mediated by a cellular liver transfer protein that is specific for the RRR form of α-tocopherol. It maintains the plasma level by selectively choosing the RRR form and recycling it into plasma lipoproteins for distribution of the vitamin to every tissue and organ in the body. The α-tocopherol transfer protein is so efficient in preserving circulating α-tocopherol that vitamin E deficiency is a rarity among humans except in diseases of malabsorption, lipoprotein abnormalities, or mutations in the gene responsible for the synthesis of the α-tocopherol transfer protein.

Deficiency. When deficiencies of the vitamin occur in humans as a consequence of acquired malabsorption of genetic abnormalities of lipoproteins or of the transfer protein, the major symptoms that develop are in the nervous system. Ataxia (lack of muscular coordination) and other neurologic symptoms result in severe incoordination and subsequent musculoskeletal changes. Fortunately, adequate supplementation with huge doses of vitamin E, when used early enough, can prevent the neurologic syndrome from developing, or can at least stabilize the neuropathology that has developed and prevent further deterioration.

Dietary requirements. The recommended daily allowance for vitamin E is 15 milligrams, which is present in the usual Western diet. Reports of studies with supplemental vitamin E are confusing because they do not specify whether natural or synthetic α-tocopherol is employed and the dose used is frequently in International Units (IU). (Since IU were originally derived from biological testing in rodents, a weight basis in milligrams is more precise.) Although it has been traditionally accepted that the biologic effective relation between natural and synthetic vitamin E is 1.36:1 (and 1 mg RRR α-tocopherol equals 1.49 IU), the true ratio is more likely to be 2:1.

The rapeutic applications. Some of the pathology of a number of degenerative brain diseases have been ascribed to oxidative stress in the central nervous system. These include Alzheimer's disease. Down syndrome, and perhaps even Parkinson's disease and Friedreich's ataxia. The possibility of favorably affecting the development and course of these diseases has made supplemental vitamin E therapy an attractive choice. Studies of treatment programs of 150 mg of the natural-source vitamin E show higher stable plasma levels without any signs of toxicity. *See* ALZHEIMER'S DISEASE; DOWN SYNDROME; PARKINSON'S DISEASE; VITAMIN. Herbert J. Kayden

Bibliography. *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*, Food and Nutrition Board, Institute of Medicine, National Academy Press, Washington DC, 2000; H. J. Kayden and M. G. Traber, Absorption, lipoprotein transport, and regulation of plasma concentrations of vitamin E in humans, *J. Lipid Res.*, 34:343, 1993.

Vitamin K

A group of compounds derived from 2-methyl-1,4-naphthoquinone that prevent bleeding in mammals and birds. Vitamin K_1 (phylloquinone) is produced by green plants; a related form, vitamin K_2 (menaquinone), is produced by intestinal bacteria. Chemically synthesized forms include vitamin K_1, K_2, and menadione (vitamin K_3). All of these compounds are fat-soluble liquids at room temperature that become biologically inactive when exposed to light or alkali. *See* VITAMIN.

Physiological activities. Vitamin K_1 (phylloquinone) is a photosynthesis cofactor in plants, and

various forms of vitamin K_2 (menaquinones) participate in energy transfer reactions in bacteria (**illus.** *a*). Mammals and many other animals need vitamin K hydroquinone (the active form) as a cofactor for the specific synthesis of the amino acid γ-carboxyglutamate (Gla) in certain proteins, which enable the proteins to bind calcium and phospholipids with high specificity. The synthesis of one Gla residue generates one vitamin K 2,3-epoxide molecule, which may be reduced again to its hydroquinone form for the next reaction (see illus.). Gla is an essential part of coagulation factors and of other proteins that regulate blood clotting (hence the disruption of blood clotting and internal bleeding associated with vitamin K deficiency). Two Gla-containing proteins (osteocalcin and matrix Gla protein) control mineralization of bone and arterial walls; others, such as gas6, influence cell growth. Gla proteins have also been found in corals and in the toxins of certain marine snails; however, their exact role in these organisms has not been defined. *See* AMINO ACIDS; BLOOD; COENZYME; PROTEIN.

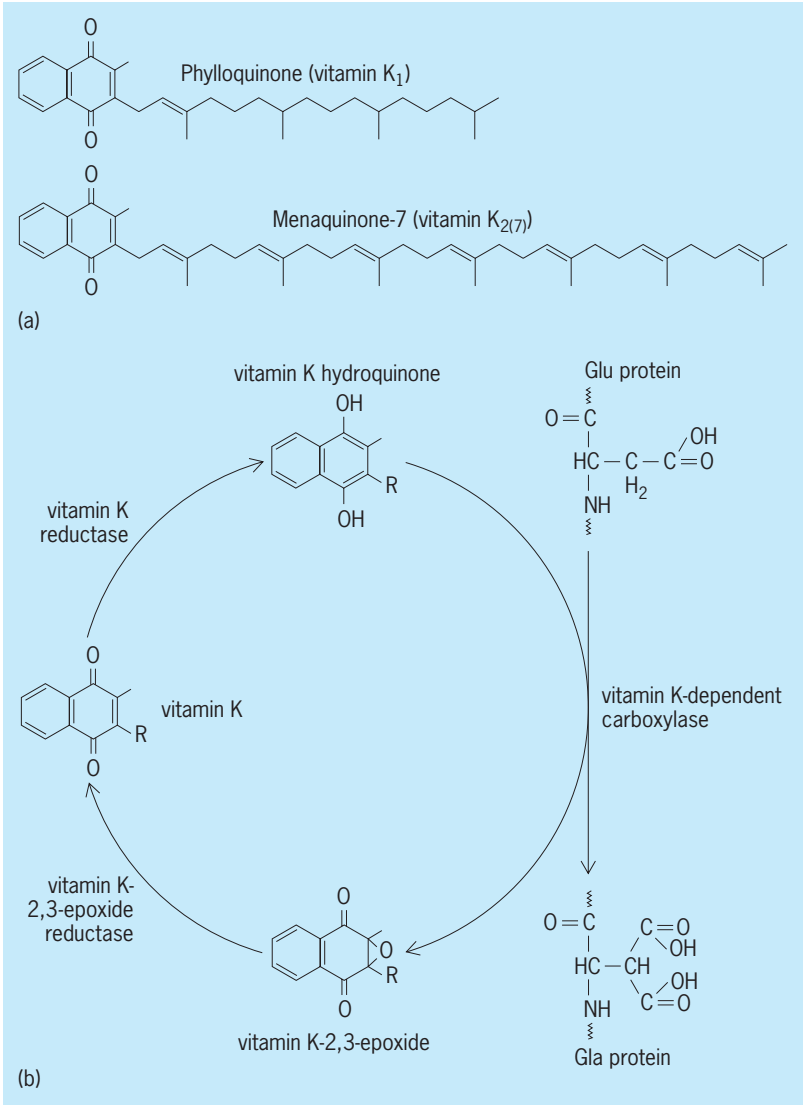
Analysis. Vitamin K in biological samples can be measured with great specificity by high-performance liquid chromatography (HPLC) or gas chromatography–mass spectrometry (GC-MS), usually after extraction and prepurification. Because the concentrations in most biological samples are low (1–500 picomoles per liter in human tissue), detection methods have to be very sensitive. Vitamin K sufficiency can be assessed with immunologic and other assays that detect the lack of Gla residues in specific vitamin K–dependent proteins such as prothrombin and osteocalcin. *See* CHROMATOGRAPHY; IMMUNOASSAY.

Requirements and deficiency. Humans depend on continuous vitamin K supplies, since storage is minimal. Good dietary sources of vitamin K_1 are green vegetables and fruits; certain fermented Asian foods, especially natto, have a high vitamin K_2 content. There is no indication that the consumption of large amounts of vitamin K, especially in the form of natural foods, has unfavorable effects on health.

Intestinal bacteria produce vitamin K_2, and under most circumstances enough is absorbed to prevent bleeding. However, spontaneous bleeding occurs if both dietary intake and production by intestinal bacteria are persistently low. In many countries, newborn infants routinely receive one-time injections of vitamin K to reduce their risk of cerebral hemorrhage. Other health risks related to inadequate vitamin K intake may include accelerated loss of bone minerals and hardening of arteries.

Antagonists. Various natural and synthetic 4-hydroxycoumarins are potent vitamin K antagonists that inhibit regeneration of the active hydroquinone form. Vitamin K antagonists are commonly used in medicine to prevent thrombosis related to operations, damaged heart valves, or abnormal heart rhythm. The antagonists have to be individually dosed, and people using this type of medication must keep their vitamin K intake constant. Antagonists are also used as a slow-acting rodent poison. The effects of accidental or intentional poisoning can be blunted by administration of vitamin K. *See* ENZYME INHIBITION; THROMBOSIS.

Martin Kohlmeier

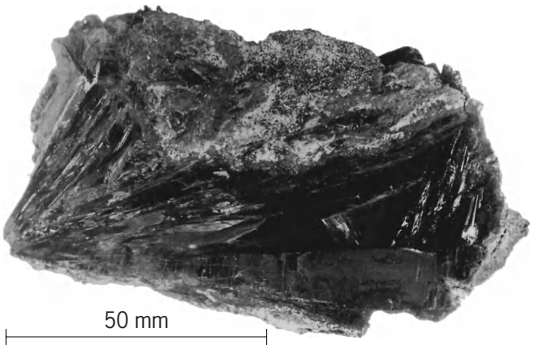


Vitamin K. (*a*) Natural forms. (*b*) Regeneration of the hydroquinone for γ-carboxyglutamate (Gla) synthesis; Glu = glutamic acid. Vitamin K functions by carboxylating certain glutamate residues in a few proteins. With each carboxylation reaction, a vitamin K molecule becomes oxidized. A microsomal enzyme system in liver, bone, and other tissues regenerates the active form, vitamin K hydroquinone. Phylloquinone and the menaquinones from bacteria work equally well in these reactions. They differ only in length and degree of saturation of their side chains; the redox-active ring system is the same.

Bibliography. J. J. Lipsky, Nutritional sources of vitamin K, *Mayo Clinic Proc.*, 69:462–466, 1994; M. J. Shearer, Vitamin K, *Lancet*, 345:229–234, 1995; M. J. Shearer, A. Bach, and M. Kohlmeier, Chemistry, nutritional sources, tissue distribution and metabolism of vitamin K with special reference to bone health, *J Nutr.*, 126:1181S–1186S, 1996.

Vivianite

A hydrated ferrous phosphate mineral, with the formula $Fe_3(PO_4)_2 \cdot 8H_2O$, for which the vivianite group is named. Other important members of this



Vivianite from Wannon River, Victoria, Australia. (*American Museum of Natural History specimen*)

group are annabergite and erythrite.

Usually ferric iron is present as the result of oxidation, although this does not change the crystal structure appreciably. It crystallizes in the monoclinic system, with crystals generally prismatic (see **illus.**). Vivianite also occurs in earthy form and as globular and encrusting masses of fibrous structure. Crystals are colorless and transparent when fresh. Oxidation changes the color progressively to pale blue, greenish blue, dark blue, or bluish black.

Vivianite is widespread as a secondary mineral in gossans of metallic ore deposits, in weathered Mn-Fe phosphatic pegmatites, and in alluvial and sedimentary deposits associated with bone and other organic remains. As odontolite in fossil bone or teeth, it is often mistaken for turquois.

Erythrite and annabergite are hydrated arsenates of cobalt and nickel, $(Co,Ni)_3(AsO_4)_2{\cdot}8H_2O$. Cobalt and nickel substitute mutually to form a complete series between the end members $Co_3(AsO_4)_2{\cdot}8H_2O$ and $Ni_3(AsO_4)_2{\cdot}8H_2O$. Erythrite includes the half of the series with Co > Ni, while annabergite includes the half with Ni > Co. Both minerals crystallize in the monoclinic system. Crystals, although rare and of poor quality, are prismatic to acicular, and transparent to translucent. Erythrite commonly occurs as radial groups, globular shapes, and earth masses. Annabergite develops as finely crystalline crusts and as earthy masses. Erythrite is crimson red and peach red. Increasing the Ni content changes the color progressively to pale pink, pale green, and apple green in annabergite.

These secondary minerals develop during oxidation of cobalt and nickel arsenides. Erythrite is the more common; both occur in Europe and North America.

Wayne R. Lowell

Vocal cords

The pair of elastic, fibered bands inside the human larynx. The cords are covered with a mucous membrane and pass horizontally backward from the thyroid cartilage (Adam's apple) to insert on the smaller, paired arytenoid cartilages at the back of the larynx. The vocal cords act as sphincters for air regulation and may be vibrated to produce sounds. Separation, approximation, and alteration of tension are produced by action of laryngeal muscles acting on the pivoting arytenoids. Innervation is through branches of the vagus nerve. Vibration of the cords produces fundamental sounds and overtones. These can be modified by the strength of the air current, the size and shape of the glottis (the opening between the cords), and tension in the cords. *See* LARYNX.

Among mammals only humans produce an elaborate articulate speech, although the basic structures for sound production are similar in all and rudiments of speech exist in many primates and other mammals. Sound in birds is produced in the syrinx, located at the ventral end of the trachea; many species are excellent mimics and some, such as the parrots, can achieve a high degree of speechlike vocalization. *See* ANIMAL COMMUNICATION; SPEECH.

Walter Bock

Voice over IP

A technology that transports voice using data packets instead of circuit switching, which the traditional public switched telephone network (PSTN) uses. Voice over IP (VoIP), using packet technology, allows for more efficient transport of voice while providing the quality of service and reliability of PSTN.

Comparison of VoIP and PSTN. The present public switched telephone network transfers voice by converting speech received from the telephone into 64-kbps (kilobits per second) digital data and transporting it in a timeslot (limited time interval) that is periodically inserted in a higher-capacity signal, a procedure referred to as time-division multiplexing (TDM). Two timeslots, one for each direction, are allocated for each phone call. These timeslots are set up by the signaling function of the public switched telephone network and kept in use for the duration of the call. The timeslots are switched in and out of use based on the calls taking place at any point in time, and are referred to as being circuit-switched. *See* TELEPHONE SERVICE.

Although the public switched telephone network can carry nonvoice data, it does so with a modem. The modem fits into the analog bandwidth of a standard telephone, allowing it to fit into the same two 64-kbps timeslots required by voice. Again these timeslots are reserved for use during the length of the call and required even when an end user does not have data to send or receive, a procedure which is not very efficient for data service, which is typically asymmetrical and bursty.

By contrast, voice over IP uses IP packets to carry speech over a data network. Only those packets that contain speech need to be transported, thereby allowing voice over IP to improve bandwidth efficiency by transporting packets only in the direction of the call participant who is listening. To further improve efficiency, voice over IP uses speech-compression algorithms to reduce speech from 64 to 2.4–8 kbps. This helps to offset the additional bytes of information (overhead) required by the packet headers, which contain the information needed to route, check, and reassemble the packets

(**Fig. 1**). These packets are transmitted every 10 milliseconds, resulting in a bandwidth of 48 kbps during speech. *See* DATA COMMUNICATIONS; PACKET SWITCHING.

Unlike the public switched telephone network using time-division multiplexing, the IP network does not allocate specific timeslots for a particular user, although newer technology does make it possible to guarantee bandwidth over the network. This allows the IP network to take advantage of silent periods of a normal call by not sending any packets, further increasing its efficiency.

While the public switched telephone network was developed for voice and backfitted for data, IP was developed for data and is being backfitted for voice. The ubiquity of IP allows the convergence of new services dependent on voice and data. Voice over IP is being expanded to support applications such as videophone, video conferencing, and whiteboard conferencing (teleconferencing involving the use of whiteboards, which are areas on display screens on which multiple users can write or draw). *See* TELECONFERENCING; VIDEOTELEPHONY.

VoIP-PSTN interoperation. Voice over IP can take place over any portion of the transmission path of a particular call, even over the public switched telephone network. In its original instantiation, voice over IP used the public switched telephone network by compressing speech, placing it into IP packets, and transporting it from one computer to another using a modem signal. Today service providers are using packet technology in addition to the public switched telephone network for transmitting voice, using gateways to provide the interface between circuit-switched and packet-switched technology.

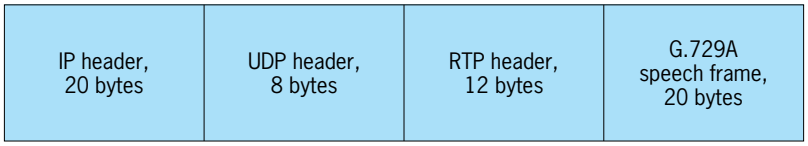


Fig. 1. Internet Protocol (IP) packet with voice over IP speech frame (encoded data) and headers. UDP = user datagram protocol. RTP = real-time protocol. G.729A is an audio data compression algorithm for voice.

The most critical development has been in technology that allows the public switched telephone network to communicate with the IP network, transforming an IP call to a public switched telephone network call, and vice versa. **Figure 2** is an example of a network that combines voice over IP and the public switched telephone network. In this example, a standard analog phone user, caller A, calls an IP phone user, caller B. Once the number is dialed, the switch processing the number uses the Signaling System 7 (SS7) protocol to send this request to the signaling gateway, which converts the SS7 messages into a Session Initiation Protocol (SIP) message. The call proceeds to caller B, who answers the call on either an SIP telephone or a personal computer running SIP client software. Once the media gateway controller detects that caller B picked up the telephone, it enables the packet voice gateway. Caller B's telephony equipment (either a voice over IP phone or a personal computer) negotiates capabilities between itself and the media gateway in the form of security and speech coding algorithms. At the completion of these negotiations, the connection between the media gateway and caller B is completed and voice is carried between callers A and B. The packet voice gateway or media gateway converts time-division multiplexing used on the public

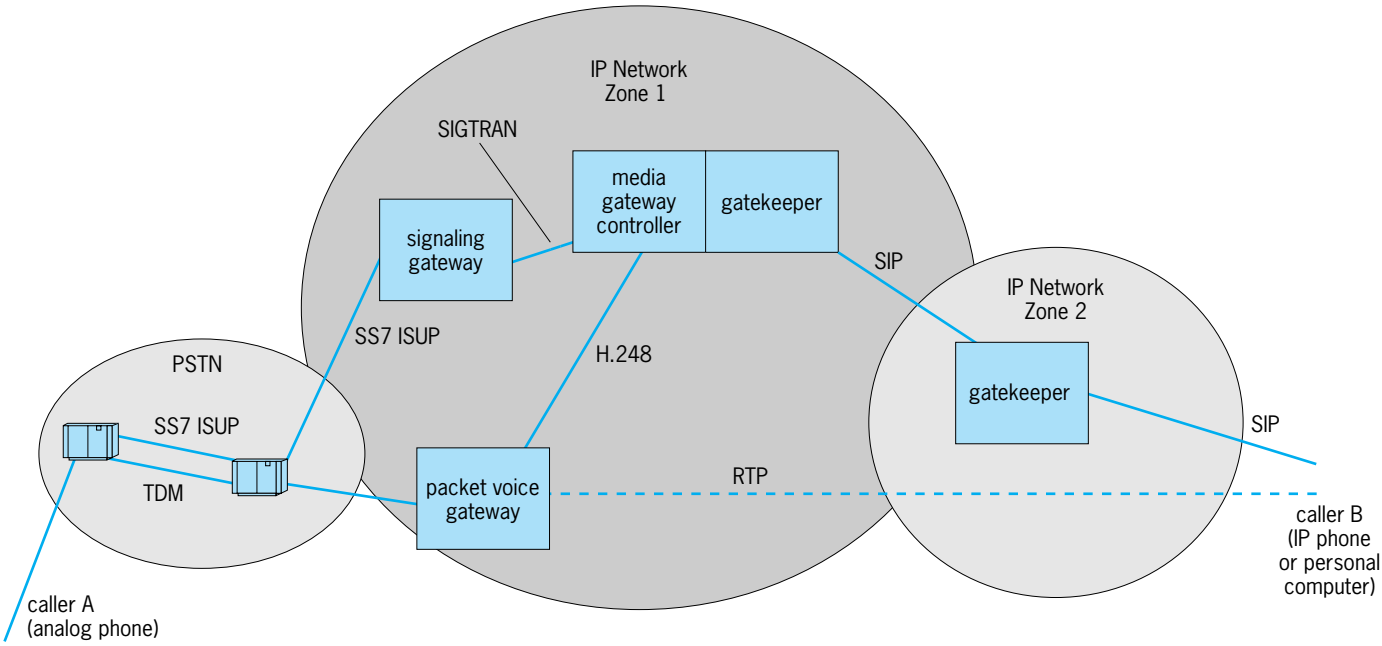


Fig. 2. Example of network operation between voice over IP and the public switched telephone network (PSTN). SS7 = Signaling System 7. ISUP = ISDN User Part. TDM = time-division multiplexing. SIGTRAN = Signaling Transport. RTP = real-time protocol. SIP = Session Initiation Protocol.

switched telephone network to and from voice being carried in real-time protocol (RTP) packets in the Internet network. *See* INTERNET.

This example happens to use SIP for the voice over Internet Protocol within the Internet network, but either the Media Gateway Control Protocol (MGCP) or H.323 might also be used. The Media Gateway Control Protocol attempts to look very similar to the public switched telephone network by adding services and control information to a central network, allowing the endpoints to be dumb. Both H.323 and SIP use a distributed architecture with intelligent endpoints and call control elements such as signaling gateways and gatekeepers, which help make sure that a call is routed to the appropriate destination. The distributed model provides for increased flexibility of applications, while the centralized approach provides an easier management of call control and provisioning (programming or instructing equipment to operate in a desired way).

Other common examples for voice over IP network configurations are end-to-end, business or enterprise services, and trunking. End-to-end, which was the initial application for voice over IP, is typically associated with low-quality calls, as the calls often go over the public switched telephone network with the end users being bandwidth-constrained. Business or enterprise applications are associated with advanced services, and their operation over high-speed local-area networks provides plenty of bandwidth. Trunking is the most recent application for voice over IP and allows service providers to transport efficiently both voice and data over their networks.

Security. Most people think of the public switched telephone network as being secure (which it is, relatively speaking), since the network resources are dedicated for the duration of a call. Packet networks or the Internet are much less secure. They are more often compared to party lines, where a piece of equipment is told to listen only for packets intended for it, but can operate in a promiscuous mode where all packets on the network can be received. This is undesirable for end users, but even more problematic is the ease with which hackers have been able to attack IP networks.

Internet Protocol networks have been broken into worldwide, and in numerous cases caused equipment to be taken off line or rendered temporarily useless. In order to build reliability into a voice over IP network, service providers must be able to increase its security level to the level of the public switched telephone network. Numerous tools and protocols have been developed for helping to secure IP networks such as firewalls, intrusion detection and prevention elements, virus scanners, and virtual private network (VPN) appliances. IP Security (IPSec), Secure Sockets Layer (SSL), and voice over IP–specific protocols are being developed, but new advances in voice over Internet Protocols and applications constantly pose new challenges to IP networks that must repeatedly upgrade their security measures to prevent network outages.

Quality of service. The public switched telephone network based on time-division multiplexing has fixed delays, while IP networks have variable delays based on the path that the data take through the network and the degree of congestion of the network. These issues can disrupt the voice by causing packets to arrive out of order or be dropped. Voice over IP software is designed to make adjustments for out-of-order packets, and speech algorithms help to adjust for lost packets, but they cannot overcome these impairments by themselves.

When delays become too great, the normal flow of a conversation is disrupted. For example, when one party finishes talking there will be an inordinate amount of time between the end of speech and a response from the far end. To address this issue, protocols [such as Multiprotocol Label Switching (MPLS) and differentiated services] are being implemented to allow for the transport of delay-sensitive data to improve performance. These protocols use information contained within packet headers to help prioritize packets so that those carrying speech (or other delay-sensitive traffic) can be routed prior to other traffic. In the case of MPLS, more overhead is added to the packet header.

Companies are also starting to increase the available bandwidth for Internet traffic and limiting the amount of traffic that can go across a particular link to improve real-time performance. When few people are trying to access or use any link, the delay is small enough not to be noticeable; but when more data go across the network, delays increase to the point where call quality is unacceptable. Home users are improving performance by using cable modems and digital subscriber lines (DSL) to increase the available bandwidth for voice over IP.

Cost savings. The first reason for using voice over IP was the cost savings that resulted from allowing home users to bypass the charges associated with long-distance calls. Voice over IP allows service providers to bypass access fees charged for using the final mile to the home since it is presently charged as data. Ultimately, these fees may change and voice over IP may become as regulated as the present public switched telephone network, thereby reducing much of today's cost advantage.

Daniel Heer

Bibliography. K. Camp, *IP Telephone Demystified*, McGraw-Hill, 2002; D. Collins, *Carrier Grade Voice over IP*, 2d ed., McGraw-Hill, 2002; D. Minoli and E. Minoli, *Delivering Voice over IP Networks*, 2d ed., Wiley, 2002.

Voice response

The generation of synthetic speech signals in order to convey information to listeners, usually based upon a verbal or textual request by the users. This speech synthesis typically employs a computer program and requires access to storage of portions of speech previously spoken by humans. The naturalness of the synthetic voice depends on several factors, including the vocabulary of words to

pronounce, the amount of stored speech, and the complexity of the synthesis programs. The most basic voice response simply plays back appropriate short verbal responses, which are only copies of human speech signals stored using digital sampling technology. The most universal systems, on the other hand, are capable of transforming any given text into comprehensible speech for a given language. These latter systems so far exist for only 20 or so of the world's major languages, and are flawed in producing speech that, while usually intelligible, sounds unnatural.

Voice response is also known as text-to-speech synthesis (TTS) because the task usually has as input a textual message (to be spoken by the machine). The text could be in tabular form (for example, reading aloud a set of numbers), or, more typically, formatted as normal sentences. Speech synthesizers are much more flexible and universal than their speech-recognition counterparts, for which human talkers must significantly constrain their verbal input to the machines in order to achieve accurate recognition. In TTS, a computer database usually determines the text to be synthetically spoken, following an automatic analysis of each user request. The user may pose the request in response to a menu of inquiries (for example, by an automated telephone dialogue, by pushing a sequence of handset keys, or by a series of brief verbal responses). Thus, the term "voice response" is used to describe the synthetic speech as an output to a user inquiry. The value of such a synthetic voice is the capability of efficiently receiving information from a computer without needing a computer screen or printer. Given the prevalence of telephones, as well as the difficulty of reading small computer screens on many portable computer devices, voice response is a convenient way to get data. *See* SPEECH RECOGNITION.

Capabilities. Voice response systems have advanced significantly in recent decades. Mechanical synthesizers approximating the actual airflow in the human vocal tract have existed, at least in primitive form, for centuries. However, the advent of practical computers in the 1960s spurred the development of much better synthetic speech. Prior to the introduction of special-purpose digital signal processing (DSP) chips in the late 1970s, synthetic speech was limited to large computers. Nowadays, voice response is becoming widespread on inexpensive devices.

While TTS products vary significantly in quality and cost, most produce generally intelligible speech, but all suffer some failings in achieving naturalness. They often sound as if they have a significant (nonhuman, machine) accent. A machine that can talk with proficiency equal to that of native speakers (that is, synthetic speech which listeners cannot distinguish from human speech) is still far from realization. Natural speech output is feasible if the desired vocabulary is very limited, and the system merely concatenates lengthy stored speech units (that is, outputs a sequence of previously spoken words or phrases, stored as coded speech). Thus, complete TTS systems which accept any input text in a chosen language (including new words as well as typographical errors) must be distinguished from more primitive voice-response systems of very limited vocabulary.

Design trade-offs. Major advances in commercial synthesizers are due to improvements in computer technology and synthesis methodology. The design of TTS trades off conflicting demands: maximum speech quality, but minimum memory space, algorithmic complexity, and computational speed. Inferior sound quality is usually due to inadequate modeling of three aspects of human speech production: coarticulation, intonation, and vocal-tract excitation.

Minimum-complexity, large-memory systems. The simplest approach to voice response is to digitally sample natural speech and output the samples later as needed. While compact disks store audio signals using 16-bit sampling at 44,100 samples per second, such high bit rates are rarely used for TTS, due to cost. A more common Nyquist sampling rate is 10,000 samples per second, which still preserves sound frequencies up to almost 5 kHz, allowing quite natural speech. High-frequency energy in fricative sounds is severely attenuated (but less so than on telephone lines), but this usually has little impact on intelligibility. Straightforward sampling [that is, linear pulse-code modulation (PCM)] requires 12 bits per sample, which requires memory at 120 kbits/s. Such high data rates are prohibitive except for applications with very small vocabularies. Even in cases with more limited bandwidth (for example, 8000 samples per second in telephone applications) and more advanced coding schemes [ranging from simple logarithmic coding at 64 kbits/s to code-excited linear prediction (CELP) at 4–10 kbits/s], the straightforward playback approach is unacceptable for general TTS. Despite rapidly decreasing costs for computer memory, it will remain impossible to store all the necessary speech signals except for applications with very restricted vocabulary needs. *See* COMPACT DISK; DATA COMPRESSION; INFORMATION THEORY; PULSE MODULATION.

Generally, voice response requires the conversion of any input text into a speech waveform via algorithms which transform previously coded speech. Such speech synthesis can be characterized by the size and type of speech units concatenated to yield the output, as well as by the method used to code, store, and synthesize. If entire phrases or sentences serve as units, output quality can be very high, but a well-chosen speaker must pronounce each phrase (during a system development stage) with timing and intonation appropriate for all sentences in which it could later be needed. Thus, if a phrase (of one or more words) could occur in several syntactic contexts, its pronunciation should be recorded with appropriate variation. Merely concatenating words which were originally spoken separately often leads to lowered output quality. The duration, spectrum, and pitch of stored units must be adjusted during concatenation since such unit features vary in sentential context (especially for smaller speech units, which need more frequent concatenations).

Minimum-storage, complex systems. A voice response system which minimizes memory needs generates synthetic speech from sequences of brief basic sounds and has great flexibility. (Like the other extreme noted above, such cases have practical applications, but most TTS lies somewhere in between.) Since most languages have only 30–40 phonemes (distinct linguistic sounds), storing units of such size and number is trivial. However, the spectral features of these short concatenated sounds (lasting 50–200 ms) must be adjusted at their frequent boundaries to avoid severely discontinuous speech. Normal pronunciation of each phoneme in an utterance depends heavily on its phonetic context (for example, on neighboring phonemes, intonation, and speaking rate). The adjustment process and the need to calculate an appropriate intonation for each context lead to complicated synthesizers with correspondingly less natural output speech.

Commercial synthesizers are primarily based on concatenation of word or phoneme units, although syllables, demisyllables, and diphones have been tried. (Diphones are phoneme-sized units, divided in the middle of each phoneme, thus preserving in each diphone the transition between adjacent phonemes.) The need to smooth spectral parameters at the boundaries between units decreases as concatenation units increase in size, because of the fewer boundaries in the synthesized speech. Such smoothing is much simpler when the joined units have similar spectra at the boundaries (as is the case with diphones). However, coarticulation (the spectral effects of adjacent sounds on each other) often extends over several phonemes. For example, in "strew," a speaker's lips will round in anticipation of the vowel during the preceding three consonants, /s/, /t/, and /r/, with corresponding acoustic effects which short diphone units cannot account for.

Storage of varied speech units. Current synthesizers usually compromise between the extremes of minimizing storage and complexity. One approach is to store thousands of speech units of varying size, which can be automatically extracted from natural speech. In contrast to automatic speech recognition, where segmentation of speech into pertinent units is very difficult, TTS training exploits prior knowledge of the text (the training speaker reads a furnished text).

The modifications needed to concatenate speech units for synthesis are similar to those of other speech applications such as time expansion and compression, which involve the slowing down or acceleration of stored utterances. Such modifications can be as simple as cut-and-splice, where individual pitch periods are duplicated or deleted periodically. (A pitch period is the time interval between successive vocal-cord or vocal-fold closures, 4–15 ms.) For more natural-sounding time-scale and pitch-scale modifications, more elaborate methods using discrete Fourier transforms or sine-wave models can be employed. *See* FOURIER SERIES AND TRANSFORMS.

Memory size. English has over 300,000 words, although only 50,000 can be considered common (and most people use only about 5000). These words can be generated more efficiently using about 12,000 morphemes, the basic meaningful elements that make up words. For example, the word "antidisestablishmentarianism" consists of the root morpheme "establish" plus two prefix and four suffix morphemes. Approximately 80,000 vectors of spectral parameters (each representing the acoustical output of a vocal tract shape) would be needed for a morpheme memory, whereas a word memory could exceed 1,000,000 entries. If the speech unit is smaller than a syllable, little memory savings are gained by limiting vocabulary; 2000 diphones can generate many thousands of words, but a much smaller vocabulary of 1000 words would still require 1000 diphones.

For large-memory TTS, a serious issue is speaker fatigue, since each speaker whose voice is simulated must utter all speech units as uniformly as possible. Speakers usually find this very arduous for more than a few thousand short phrases. Furthermore, the longer the time span of recording, the more likely is nonuniformity to be present, and the resulting uneven units lead to rough-sounding speech.

Synthesis method. Most voice-response systems are terminal-analog synthesizers, meaning that they model the speech output after filtering by the vocal tract, without explicitly accounting for articulator movements. (Alternative articulatory synthesizers have not been practical so far, due to the difficulty of obtaining accurate three-dimensional vocal-tract representations and of modeling the system with a small set of parameters.) Until recently, waveform methods for synthesis, were used only with small vocabularies (for example, a few minutes of speech). For greater flexibility, parametric methods are necessary and have dominated the field. Such synthesis, however, has often been limited in speech quality owing to inadequate modeling of natural speech production. A standard speech coding method called linear predictive coding (LPC) has often been used to store speech units efficiently and play them back. This suffices for many simple applications (for example, telephone directory assistance). However, if a standard voice and small vocabulary are insufficient, the synthesizer manufacturer must process speech spoken by specific users to establish a custom vocabulary. Such as hoc processing can yield very compact representations, but is expensive.

Unrestricted-text (TTS) systems. Synthesizers that accept general text as input need a linguistic processor to convert the text into phonetic symbols in order to access the appropriate stored speech units. One task is to convert letters into phonemes. This may be as simple as a table look-up: a computer dictionary with an entry for each word in the chosen language, noting its pronunciation (including syllable stress), syntactic category, and possibly some semantic information. Many systems also have language-dependent rules, which examine the context of each letter in a word to determine how it is pronounced; for example, the letter [p] in English is pronounced /p/, except before the letter [h] (for example, in "telephone"; however, it has normal pronunciation in "cupholder"). English needs hundreds of such rules.

TTS often employs these rules as a back-up procedure to handle new words, foreign words, and typographical mistakes (that is, cases not in the dictionary). *See* PHONETICS.

Languages in which spelling follows phonetics more closely (for example, Spanish) can be modeled with very few rules since each letter has normally only one pronunciation. (The same holds for character-based languages such as Chinese, but those alphabets are much larger than 26 letters.) Letter-to-phoneme rules developed manually for many languages are capable of high precision, especially when combined with a dictionary to handle exceptional cases. Errors in phonetic transcription with such advanced systems are almost always due to proper nouns (such as names) or foreign words. Such words may obey rules different from those of the modeled language and are often capitalized (or italicized) in normal text (and can thus be automatically identified for special processing).

Letter-to-phoneme rule sets produced via neural networks have been much less successful than those developed manually. Neural nets have found greater success in speech recognition, where the number of possible utterances and acoustic variation is enormous, compared to the number of words in a language. *See* NEURAL NETWORK.

Some systems employ a word decomposition algorithm, which attempts to strip prefixes and suffixes from each word in an input text. Since there are only a few dozen such affixes in most languages and since they can affect pronunciation (for example, the third vowel in "algebra" versus "algebraic"), such a decomposition procedure can reduce memory requirements (at the expense of extra computation). On the other hand, syntactic (and less often, semantic) information is easily stored in a dictionary. Combined with a natural language parser to determine linguistic structures in the input text, this allows artificial specification of intonation by rule, that is, locating pitch and duration effects to simulate natural intonation. Many synthesizers still forego parsers as too complex or unreliable, and use simplistic intonation rules. Poor handling of intonation is a major reason why much TTS sounds unnatural.

Formant synthesis. Traditionally, TTS has employed a cascade or parallel structure of digital filters, each simulating one resonance (formant) in the vocal tract being simulated. The filters' excitation is a regularly spaced sequence of impulses (for periodic voiced speech) and pseudo-random noise (for unvoiced speech). The cascade structure approximates speech spectra well for vowels and allows simple control with one amplitude parameter. The lowest four formant frequencies and three corresponding bandwidths vary as a function of time. (Variation in higher frequencies has little perceptual effect.) *See* DIGITAL FILTER.

Linear predictive coding (LPC) synthesis. TTS has often avoided the need to manually develope coarticulation rules (for formant synthesis), instead using the automatic method of LPC. The filter for LPC has a simpler structure than that for formant synthesis becuase all spectral properties of speech (except for intensity and periodicity, which present equal challenges to most TTS methods) are included in the LPC spectral parameters (coefficients). A lattice filter is often used because the multiplier coefficients (modeling boundaries between adjacent cylindrical sections of a vocal tract) can be linearly interpolated between successive frames (for smoother speech) without yielding an unstable filter. (A frame is the length of time over which speech analysis is performed, 10–30 ms.) Formant synthesis is more flexible than LPC in allowing simple transformations to simulate different voices. Modifying formant frequencies can alter speaker-related aspects of voice much more easily than with LPC. *See* ELECTRIC FILTER.

Waveform concatenation. A more recent TTS alternative is waveform synthesis, which can yield very high quality speech at the cost of increased memory and decreased flexibility. In the PSOLA (pitch-synchronous overlap-and-add) method, for example, brief speech waveform units are concatenated, effectively fading out one unit while fading in the next, typically with synchronized pitch periods from adjacent units. As with diphone synthesis methods, PSOLA is not readily modified to simulate voices other than that of the training speaker; for many TTS applications, one voice suffices. Early PSOLA methods suffered from a need to manually segment pitch periods in the stored speech units, problems in smoothing at unit boundaries, and a requirement for large memory (for example, 80 kbits/s). Modifications have lessened some drawbacks, while retaining freedom from intensive computation at synthesis time. For example, the MBROLA (multiband resynthesis) method resynthesizes (once) the entire diphone database at a constant average pitch period, which allows simple linear time interpolation at synthesis time and smoother spectral matching.

Synthesis of intonation. The problem of determining an appropriate intonation for each input text continues to confound TTS. In simple voice response, the stored units are large (for example, phrases), and pitch and intensity are usually stored explicitly with the spectral parameters or implicitly in the signals of waveform synthesizers. However, when smaller units are concatenated, the synthetic speech sounds unnatural unless the intonation is adjusted for context. Intonation varies significantly among languages. Although automatic statistical methods show some promise, intonation analysis has mostly been manual.

Three prosodic parameters contribute to intonation: pitch, duration, and intensity. Many intensity variations are phoneme-dependent, and stressed syllables are more intense. Pitch (due to the vibration rate of the vocal cords) is the most problematic of the intonation parameters, owing to its significant variation, both at a frame-by-frame level and globally across the sentence. In English, lexically stressed syllables of emphasized words cause pitch changes to cue the boundaries of syntactic phrases. Pitch usually follows a basic falling pattern, with superimposed excursions above and below a declination line.

Compared to pitch, duration is more tightly linked to phonemics; for example, vowels are longer than consonants, stressed syllables longer than unstressed ones, and consonants shorter in clusters. Vowels are longer when prior to voiced consonants than unvoiced ones. When speaking rate varies, vowels tend to expand or compress more than consonants do.

A major difficulty in specifying natural intonation is the lack of reliable markers in most input text to indicate intonational boundaries. Sentence-final and clause-final punctuation (. ? ! : ;) are reliable places for pauses. However, many sentences feature long sequences of words with, at most, only commas (which do not correspond well to intonation). In most European languages, a break often occurs after a (high-information) "content" word followed by a "function" word. Highlighting the final word in such sequences (with durational lengthening and a pitch rise) is often appropriate.

Different languages. Simple voice-response systems work equally well for all languages since they just play back previously stored speech units. For general TTS, however, major synthesizer components are highly language-dependent. The front end of TTS systems, dealing with letter-to-phoneme rules, the relationship between text and intonation, and different sets of phonemes, is language-dependent. The back end, representing simulation of the vocal track via digital filters, is relatively invariant across languages. Even languages with sounds (for example, clicks) other than the usual pulmonic egressives require only simple modifications.

Practical speech synthesis. Commercial synthesizers are widely available for about 10 languages. They often combine software, memory, and processing chips, and range from expensive systems providing close-to-natural speech to inexpensive personal computer programs. General digital signal processing chips are widely used for TTS. Current microprocessors can easily handle the speeds for synthesis, and indeed synthesizers exist entirely in software. Memory requirements can still be a concern, especially for some of the newer waveform concatenation systems. *See* MICROPROCESSOR.

Prospects. Speech synthesis is increasingly popular, as the cost and size of computer memory decreases. Limited-vocabulary voice response yields high quality and suffices for many applications. Increasingly, inexpensive memory has been exploited by using large inventories of speech units to overcome some coarticulation and intonation problems. These trends follow those of automatic speech recognition methods, whose stochastic methods involve very simple network models, but require massive amounts of training and memory to accommodate the large amount of variability in the way that speakers talk. While TTS does not need to model such speaker variability, it must handle well the large amount of variability (even within one speaker's voice) across the many different phonetic contexts met in normal speech. Eventually, increased understanding of how humans produce and perceive speech will yield more efficient TTS, and some combination of stochastic and knowledge-based methods will yield synthetic speech quite similar to that of humans. *See* SPEECH.

Douglas O'Shaughnessy

Bibliography. J. Allen, Overview of text-to-speech systems, in S. Furui and M. Sondhi (eds.), *Advances in Speech Signal Processing*, pp. 741–790, Marcel Dekker, New York, 1992; T. Dutoit, *An Introduction to Text-to-Speech Synthesis*, Kluwer, 1997; D. Klatt, Review of text-to-speech conversion for English, *J. Acous. Soc. Amer.*, 82:737–793, 1987.

Volatilization

The process of converting a chemical substance from a liquid or solid state to a gaseous or vapor state. Other terms used to describe the same process are vaporization, distillation, and sublimation. A substance can often be separated from another by volatilization and can then be recovered by condensation of the vapor. The substance can be made to volatilize more rapidly either by heating to increase its vapor pressure or by removal of the vapor using a stream of inert gas or a vacuum pump. Heating procedures include the volatilization of water, of mercury, or of arsenic trichloride to separate these substances from interfering elements. Chemical reactions are sometimes utilized to produce volatile products as in the release of carbon dioxide from carbonates, of ammonia in the Kjeldahl method for the determination of nitrogen, and of sulfur dioxide in the determination of sulfur in steel. Volatilization methods are generally characterized by great simplicity and ease of operation, except when high temperatures or highly corrosion-resistant materials are needed. *See* CHEMICAL SEPARATION TECHNIQUES; DISTILLATION; SUBLIMATION; VAPOR PRESSURE.

Louis Gordon; Royce W. Murray

Volcanic glass

A natural glass formed by rapid cooling of magma. Magmas typically comprise crystals and bubbles of gas within a silicate liquid. On slow cooling, the liquid portion of the magma usually crystallizes, but if cooling is sufficiently rapid, it may convert to glass—an amorphous, metastable solid that lacks the long-range microscopic order characteristic of crystalline solids. *See* LAVA.

Formation. Because it is in general easier to form glasses from viscous liquids than from highly fluid liquids, the conditions required for the formation of glasses in nature vary with the chemical composition of the liquids. For example, silica-rich, rhyolitic magmas frequently quench to glass during explosive eruptions and make up the bulk of the solid material in many pyroclastic deposits (usually as shards, pumice lumps, and other fragments); but they also can erupt quiescently to form massive glassy rocks (known as obsidian, the most common source of volcanic glass on land) even in the slowly cooled interiors of flows tens of meters thick. In contrast, more

basic, basaltic glasses (sometimes known as tachylite) are less common and rarely form in more than small quantities unless rapidly cooled in a volcanic eruption. Peles hair is an example of basaltic glass formed in this way. Because magmas cool much more rapidly when erupted under water, basaltic glasses are commonly retrieved from the quenched margins of submarine basalt flows. Small quantities of glass are frequently found in the interstices between crystals in largely crystalline rocks or as small (on the order of a few to a few hundred micrometers in diameter) inclusions trapped during growth of crystals. The low probability of nucleation in such small volumes may be a factor in the vitrification of even inviscid liquids in these inclusions. *See* OBSIDIAN; RHYOLITE.

Significance of composition. The chemical compositions of magmatic liquids are key to understanding the origin of diversity in igneous rocks. Crystalline and partially crystalline rocks are frequently formed from crystal + liquid mixtures into which crystals have been concentrated by, for example, gravitational settling or rising, or from which liquid has been preferentially extracted. It is not always possible to recognize or correct for the effects of such processes on the compositions of crystalline rocks and thereby to reconstruct actual liquid compositions. There has long been considerable interest in volcanic glasses among geologists because, provided that they are not subsequently chemically altered, these glasses can provide unequivocal information on the compositions of liquids involved in igneous processes.

Igneous petrogenesis. Detailed study of unaltered volcanic glasses has led to several important conclusions about igneous petrogenesis that were unavailable or insecure on the basis of crystalline rock studies. Examples of important results from the study of glasses from the submarine basalts of the mid-oceanic ridge environment include (1) determination of the variations in liquid chemical composition within individual provinces and at different parts of the worldwide ridge system, providing the primary constraints on the processes leading to the formation of the Earth's most abundant magma type; (2) measurement of the oxidation state of primitive magmas, and by inference, of the suboceanic mantle; and (3) measurement of the concentrations and isotopic characteristics of volatile components (for example, carbon dioxide, water, sulfur, and chlorine) in relatively undegassed magmas, providing crucial inputs into understanding of the sources and concentrations of these components in the mantle. The study of fresh obsidians and of glass inclusions in crystals in rhyolites has led to similar insights into the evolution of silicic magmas, such as the determination of the preeruptive volatile contents of rhyolitic liquids (for example, typically 4–6 wt % water, dissolved both as molecules of water and hydroxyl groups); the depths of the chambers from which such magmas erupt and the temperatures in such chambers; and the mechanisms by which relatively dry obsidian flows form (for example, by collapse of foams formed during degassing of water-rich rhyolitic magmas as they rise toward the surface). *See* IGNEOUS ROCKS; MID-OCEANIC RIDGE.

Duration. Volcanic glass is highly susceptible to alteration when it is exposed to water, even in low-temperature environments such as on or close to the Earth's surface. Basaltic glasses rapidly alter to a poorly characterized, partially crystalline material known as palagonite. Typically obsidians initially remain largely glassy upon hydration, but their compositions usually change drastically during the process, principally by exchange of oxygen, alkalies, and silica with the aqueous phase. Thicknesses of palagonite on basalt and of hydrated rinds on obsidian can, under favorable circumstances, be used to determine the dates of volcanic eruptions. Perlite, with a characteristic texture of spherical cracks, and pitchstone, defined on the basis of its luster, are the result of extensive hydration of obsidian. Hydrated glass tends to devitric (that is, crystallize) given sufficient time and even minor heating. Consequently, terrestrial glasses are rare in pre-Tertiary rocks. However, glasses that are free of water and have not been reheated have survived for several billions of years in lunar rocks and meteorites. *See* MAGMA; PITCHSTONE.

Edward M. Stolper

Volcano

A mountain or hill, generally steep-sided, formed by accumulation of magma (molten rock with associated gas and crystals) erupted through openings or volcanic vents in the Earth's crust; the term volcano also refers to the vent itself. During the evolution of a long-lived volcano, a permanent shift in the locus of principal vent activity can produce a satellitic volcanic accumulation as large as or larger than the parent volcano, in effect forming a new volcano on the flanks of the old.

Planetary exploration has revealed dramatic evidence of volcanoes and their products on the Earth's Moon, Mars, Mercury, Venus, and the moons of Jupiter (**Fig. 1**), Neptune, and Uranus on a scale much more vast than on Earth. For example, Olympus Mons, a gigantic shield volcano on Mars about 600 km (375 mi) in diameter, is larger across than the length of the Hawaiian Islands. However, only the products and landforms of terrestrial volcanic activity are described here. *See* VOLCANOLOGY; MARS; MERCURY (PLANET); MOON; NEPTUNE; URANUS; VENUS.

Volcanic vents. Volcanic vents, channelways for magma to ascend toward the surface, can be grouped into two general types: fissure and central (pipelike). Magma consolidating below the surface in fissures or pipes forms a variety of igneous bodies (**Fig. 2**), but magma breaking the surface produces fissure or pipe eruptions (**Figs. 3** and **4**). Fissures, most of them less than 10 ft (3 m) wide, may form in the summit region of a volcano, on its flanks, or near its base; central vents tend to be restricted to the summit area of a volcano. For some volcanoes or volcanic regions,

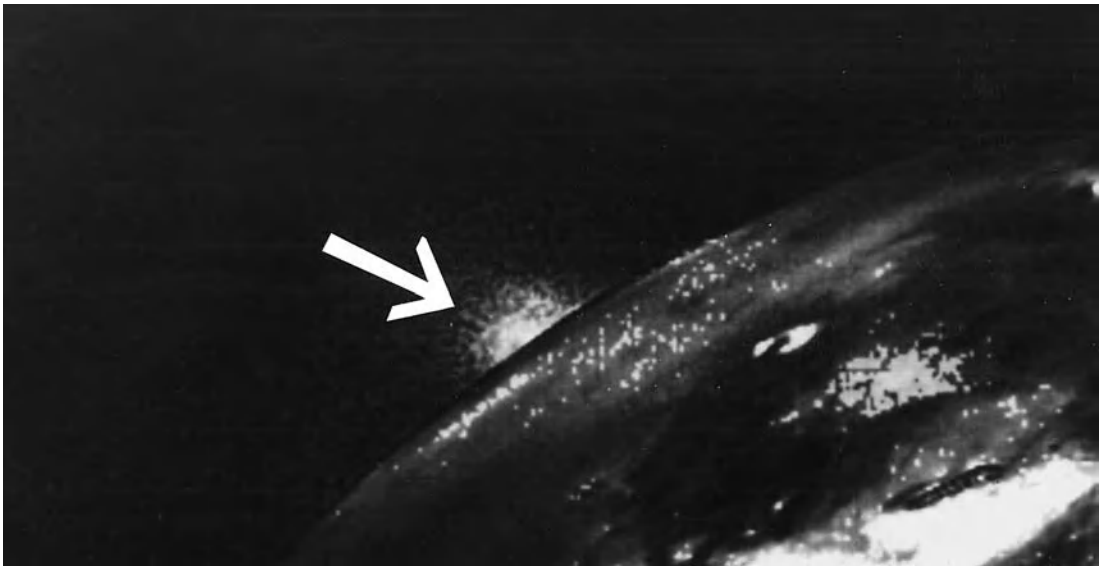


Fig. 1. Picture taken by *Voyager 2* in July 1979 showing volcanic plumes (arrow) rising approximately 60–110 mi (95–175 km) above the surface of Io, a moon of Jupiter. [In 1997 and 1999 the *Galileo* spacecraft also captured images of Io's eruptive activity.] (*NASA*)

swarms of fissure vents are clustered in swaths called rift zones.

Volcanic products. Magma erupted onto the Earth's surface is called lava. If the lava is chilled and solidifies quickly, it forms volcanic glass; slower rates of chilling result in greater crystallization before complete solidification. Lava may accrete near the vent to form various minor structures or may pour out in streams called lava flows, which may travel many tens of miles from the vents. During more violent eruption, lava torn into fragments and hurled into the air is called pyroclastic (fire-broken materials). The general term "tephra" is applied to poorly consolidated pyroclastic debris regardless of fragment or particle size. *See* CRYSTALLIZATION; LAVA; MAGMA; VOLCANIC GLASS.

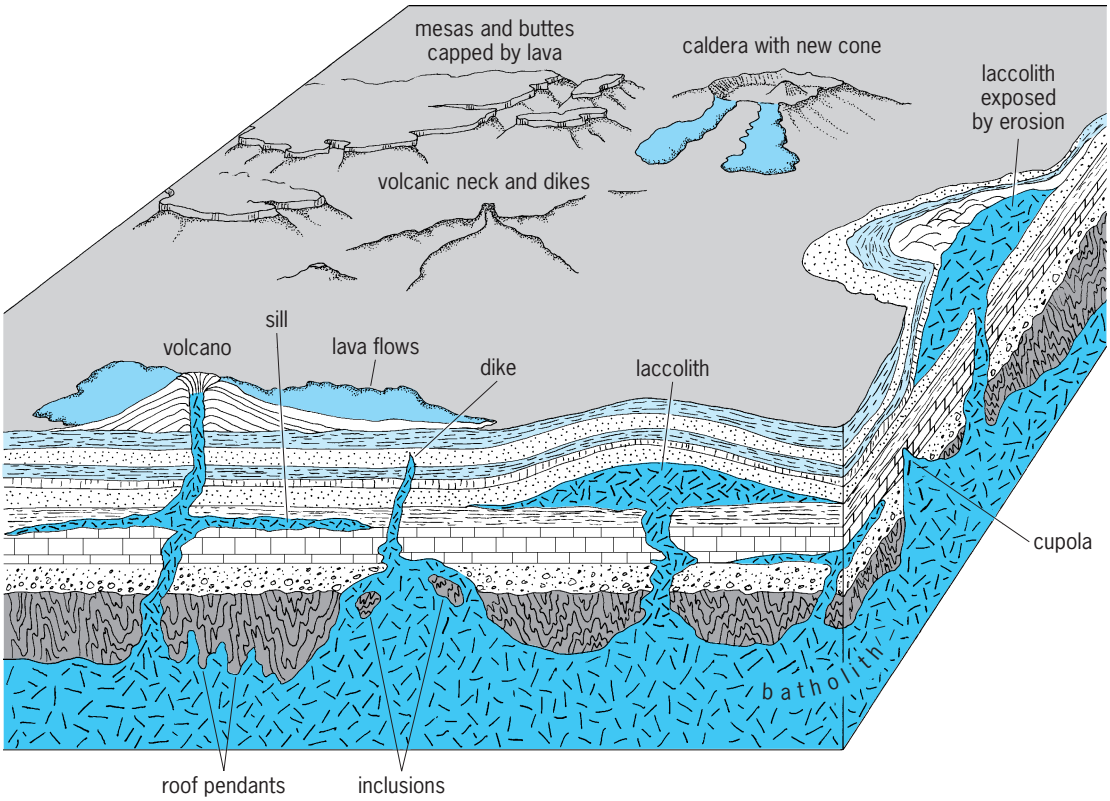


Fig. 2. Cross-sectional diagram illustrating igneous bodies in relation to geological structures and surface forms. (*After W. H. Emmons et al., eds., Geology: Principles and Processes, 5th ed., McGraw-Hill, 1960*)

The character of a volcanic eruption depends largely on the viscosity of the liquid lava; viscosity, for the purposes of this discussion, can be considered as a stickiness coefficient of the magma or lava. In general, mafic (basaltic and similar) lavas are less viscous (more fluid) than more silicic (dacitic and similar) lavas. Eruptions of mafic lavas commonly are nonexplosive or only weakly explosive and produce mostly lava flows; eruptions of silicic lavas are typically explosive and yield predominantly pyroclastic material. Some produce only pyroclastic material.

Lava flows. Lava flows are classified in terms of their surface characteristics. Flows showing smooth or hummocky, gently undulating surfaces and crusts locally wrinkled into ropelike forms are called pahoehoe. Flows showing very rough irregular surfaces covered by jagged spinose fragments resembling furnace clinker are called aa (**Fig. 5**). The terms pahoehoe and aa are of Hawaiian origin. Flows in which the fragments that constitute the upper part of the flow are fairly smooth-sided polygons are called block lava. Fluid basaltic and related mafic lavas characteristically form pahoehoe or aa flows, or flows intermediate in character between these two end members. In contrast, more viscous lavas such as andesites more commonly form block lava flows. All lava flows that have poured out on land contain various amounts of open cavities (vesicles), which mark the sites of gas bubbles formed as the rising magma reached regions of increasingly lower pressure before breaching the surface to erupt. Over geologic time, vesicles can be filled by minerals precipitated from circulating mineralizing fluids or ground water. Where pahoehoe flows enter bodies of water or wet ground, they may form heaps of irregular ellipsoids, in cross section somewhat resembling sacks of grain or pillows. Basaltic pillow lavas form an overwhelming bulk of the ocean floor. *See* ANDESITE.

Pyroclastic materials. Magma, at depth and under great pressure, contains gas in solution, but as it rises into regions of lower pressure near the surface of the Earth, the gas starts to exsolve and escape from the liquid. The gas generally escapes readily from fluid lavas, with little or no explosion; in more viscous liquids, however, the gas may acquire considerable pressure before it escapes and bursts forth in strong explosions. Thus, the emergence at the surface of highly gas-charged, viscous silicic lava may be attended by a sudden frothing as the contained gas rapidly exsolves and vesiculates. The sudden expansion of the gas may tear the froth into countless tiny shreds, each of which chills virtually instantaneously to form fragments of volcanic glass. Continued ebullition of gas results in a mass of small solid or quasisolid fragments, each surrounded by an envelope of still-expanding gas that pushes against all adjacent expanding envelopes. The net effect of this process is the isolation of the solid fragments from contact with one another, and the entire mass obtains an expansive quality that is enhanced further by the expansion of heated air trapped in the hot, moving mass. The result is a very mobile suspension of incandescent solid fragments in gas which may flow at great speed down slopes and spread out to great distances from the erupting vents, forming extensive deposits having nearly flat surfaces. When these ash flows come to rest, they commonly are still so hot that the fragments of glass stick together or even merge in the center of the deposit to form a layer of solid black obsidian. The resulting deposits are known as welded tuffs or ignimbrites; the degree of welding of ash flows is largely dependent on the

Fig. 3. Aerial view of 165-ft-high (50-m) lava fountains spurting from three en-echelon fissure vents on the northeast rift zone of Mauna Loa volcano, Hawaii, feeding a massive lava flow during its July 1975 eruption. Mauna Loa erupted again during March–April 1984, producing two 16-mi-long (25-km) flows that reached within 3–6 mi (5–10 km) of Hilo, the largest city on the island of Hawaii. (*R. T. Holcomb, USGS*)

Fig. 4. Parícutin volcano, Mexico, February 20, 1944, showing typical eruptive activity from a pipelike vent to form a cinder cone. (*Courtesy of T. Nichols*)

Fig. 5. Active pahoehoe lava flow (left) lapping against and overriding previously erupted aa lava (right) during the 1972–1974 eruption of Kilauea volcano, Hawaii. (*R. I. Tilling*, *USGS*)

temperature and thickness of the deposit. Rapidly moving incandescent ash flows have been called glowing avalanches or nuée ardentes; they can be highly destructive to life and property. For example, in 1902 a devastating nuée ardente produced during a violent eruption of Mont Pelée (island of Martinique in the Lesser Antilles) virtually destroyed the entire city of St. Pierre, killing some 30,000 persons. More recently, nuées ardentes generated during the March-April 1982 eruption of El Chichón volcano (State of Chiapas, southeastern Mexico) wiped out all settlements within 5 mi (8 km) of the volcano, resulting in the loss of more than 2000 people (**Fig. 6**). *See* OBSIDIAN; TUFF.

Pyroclastic materials are classified by the nature of the shreds and fragments: essential—the erupting molten lava; accessory—solidified lava of previous eruptions of the same volcano; and accidental—solid material of still older, not necessarily volcanic, rocks from the crust beneath the volcano. The classification of pyroclastic materials also considers size, shape, and consistency of constituent fragments and matrix. Blobs or drops of material still liquid enough to assume rounded or aerodynamically drawn-out forms during flight are known as lapilli (if 0.16–1.2 in. or 4–32 mm in average diameter) and bombs (if greater than 1.2 in. or 32 mm). Depending on their final shapes when they strike the ground, bombs are variously called cow-dung bombs, spindle or fusiform bombs, or ribbon bombs.

Fig. 6. Ruins of church (circled), the only visible remains of the village of Francisco León after being swept by nuées ardentes from the 1982 eruption of El Chichón volcano, southeastern Mexico. (*R. I. Tilling*, *USGS*)

Irregular fragments of frothy lava of bomb or lapilli size are called scoria or cinder; if the fragments are sufficiently plastic to flatten or splash as they hit, they are called spatter. The still-molten fragments of spatter often adhere to each other to form welded spatter, or agglutinate. Angular fragments larger than 1.2 in. (32 mm) either solid or too viscous to assume rounded forms during flight are known as blocks; their accumulation forms a volcanic breccia. Ejecta smaller than 0.16 in. (4 mm) are called ash, and those smaller than 0.01 in. (0.25 mm) are called dust. Indurated (hardened) volcanic ash or dust is called tuff.

The term "tephra" had been applied liberally to all pyroclastic materials, but this term should be reserved, as originally defined, for those pyroclastic deposits of air-fall origin regardless of the size of the ejected material. The individual pyroclastic fragments—the building blocks of pyroclastic rocks—are called pyroclasts. *See* PYROCLASTIC ROCKS.

Volcanic gases. In general, water vapor is the most abundant constituent in volcanic gases; the water is mostly of meteoric (atmospheric) origin, but in some volcanoes can have a significant magmatic or juvenile component. Excluding water vapor, the most abundant gases are the various species of carbon, sulfur, hydrogen, chlorine, and fluorine (such as CO_2, CO, SO_2, SO_3, H_2, H_2S, Cl_2, F_2, HCl). As discussed above, the related processes of rapid exsolution, expansion, and release of gas from the magma as it ascends near the Earth's crust provide the energy for driving and sustaining volcanic eruptions. Volcanic gases, as well as the erupted solid volcanic materials, can pose a hazard to humans. In August 1986, at Lake Nyos, a volcanic lake occupying the crater of a geologically young volcano in Cameroon (western Africa), the sudden overturn of deep lake waters caused the massive release of carbon dioxide (CO_2) gas of magmatic origin, asphyxiating more than 1700 people in low-lying areas.

Volcanic aerosols. Violent volcanic explosions may throw dust and aerosols high into the stratosphere, where it may drift across the surface of the globe for many thousands of miles. Fine ash and dust from eruptions of Icelandic volcanoes have fallen in the streets of Moscow. Studies have shown that the particles in the eruption cloud are mostly angular bits of lava, many of them glassy; but some glass spheroids also are present, as well as liquid droplets of hydrous solutions of sulfuric acid and various sulfates and chlorides. Most of the solid particles in the volcanic cloud settle out within a few days, and nearly

all settle out within a few weeks, but the gaseous aerosols (principally sulfuric acid droplets) may remain suspended in the stratosphere for several years. Such stratospheric clouds of volcanic aerosols, if sufficiently voluminous and long-lived, can have an impact on global climate. *See* ACID RAIN; AEROSOL; AIR POLLUTION.

Volcanic mudflows. Mudflows are common on steep-side volcanoes where poorly indurated or nonwelded pyroclastic material is abundant. They may form by eruptions involving the water of a crater lake; after breaching its confining walls, the water sweeps down the mountainside, incorporating and mixing with loose volcanic debris to form a slurry. Such slurries can be quite dense and commonly have a consistency similar to wet concrete. Mudflows may also form by hot or cold volcanic debris avalanches descending into streams or onto snow or ice. Probably by far the most common cause, however, is simply heavy rain saturating a thick cover of loose unstable pyroclastic material on the steep slope of the volcano, transforming the material into a mobile, water-saturated "mud," which can rush downslope at a speed as great as 50–55 mi (80–90 km) per hour. Such a dense, fast-moving mass can be highly destructive, sweeping up everything loose in its path. Volcanic mudflows can be hot or cold; they are sometimes called lahars, a term from Indonesia, where they have taken a heavy toll in property and human lives (**Fig. 7**). Fast-moving mudflows triggered by a small eruption on November 13, 1985, at Nevado del Ruiz, Colombia, killed about 25,000 people. The Ruiz catastrophe was the worst volcanic disaster in the twentieth century since the devastation associated with the eruption of Mont Pelée, Martinique in 1902.

Somewhat related to mudflows are the great floods of water, known in Iceland as Jökulhlaup, that result from rapid melting of ice by volcanic eruption beneath a glacier. A subglacial eruption in October 1996 at Grímsvötn Volcano beneath the Vatnajökull ice sheet, Iceland, produced abundant hyaloclastite. In November 1996 the catastrophic release of impounded melt water from the eruption produced a destructive glacial outburst flood that affected more than 270 mi^2 (750 km^2) and destroyed or severely damaged several bridges. *See* GLACIOLOGY.

Volcanic landforms. Much of the Earth's solid surface, on land and below the sea, has been shaped by volcanic activity. Landscape features of volcanic origin may be either positive (constructional) forms, the result of accumulation of volcanic materials, or negative forms, the result of the lack of accumulation or collapse.

Major edifices. The gross form of a volcano is largely determined by the viscosity and mode of eruption of the volcanic products, even though other factors may be operative locally on a smaller scale.

Not all volcanoes show a graceful, symmetrical cone shape, such as that exemplified by Mount Fuji, Japan, or Mayon volcano, Philippines (**Fig. 8**). In reality, most volcanoes, especially those near tectonic plate boundaries, are more irregular, though of grossly conical shape. Such volcanoes, called

Fig. 7. Houses in the Cibanjaran River buried by lahars (volcanic mudflows) generated during the 1982–1983 eruption of Galunggung volcano, Province of West Java, Indonesia. (*R. T. Holcomb, USGS*)

Fig. 8. Mayon, a symmetrical composite volcano near Legaspi, Luzon Island, Philippines. In September-October 1984, its eruption caused the evacuation of tens of thousands of residents from 36 villages. (*G. A. MacDonald*, *USGS*)

stratovolcanoes or composite volcanoes, typically erupt explosively and are composed dominantly of andesitic, relatively viscous and short lava flows, interlayered with beds of ash and cinder that thin away from the principal vents. Volcanoes constructed primarily of fluid basaltic lava flows, which may spread great distances from the vents, typically are gentle-sloped, broadly upward convex structures that resemble a Germanic warrior's shield in form. Such shield volcanoes, classic examples of which are Mauna Loa and Mauna Kea volcanoes, Hawaii (**Fig. 9**), tend to form in oceanic intraplate regions and are associated with hot-spot volcanism. The shape and size of a volcano can vary widely between the simple forms of composite and shield volcanoes, depending on magma viscosity, eruptive style (explosive versus nonexplosive), migration of vent locations, duration and complexity of eruptive history, and posteruption modifications (**Fig. 10***a*). Large shield volcanoes are much larger than the largest composite volcanoes (Fig. 10*b*).

Fig. 9. Mauna Kea volcano, 13,800 ft (4208 m) high, viewed from sea level at Hilo Bay, Hawaii. Mauna Kea shows the classic, gently sloping, broadly convex form of a shield volcano, whose surface has been peppered with numerous smaller, steeper-sided cinder cones. (*R. I. Tilling*, *USGS*)

Some of the largest volcanic edifices are not shaped like the composite or shield volcanoes. In certain regions of the world, voluminous extrusions of very fluid basaltic lava from dispersed fissure swarms have built broad, nearly flat-topped accumulations, some covering hundreds of thousands of square miles with volumes of several tens of thousands of cubic miles. These voluminous outpourings of lava are known as flood basalts or plateau basalts. Two examples in the United States are the Columbia River Plateau (**Fig. 11**) and the Snake River Plain volcanic fields; similar large features exist in India (Deccan Plateau) and South America (Paraná Basin). *See* BASALT.

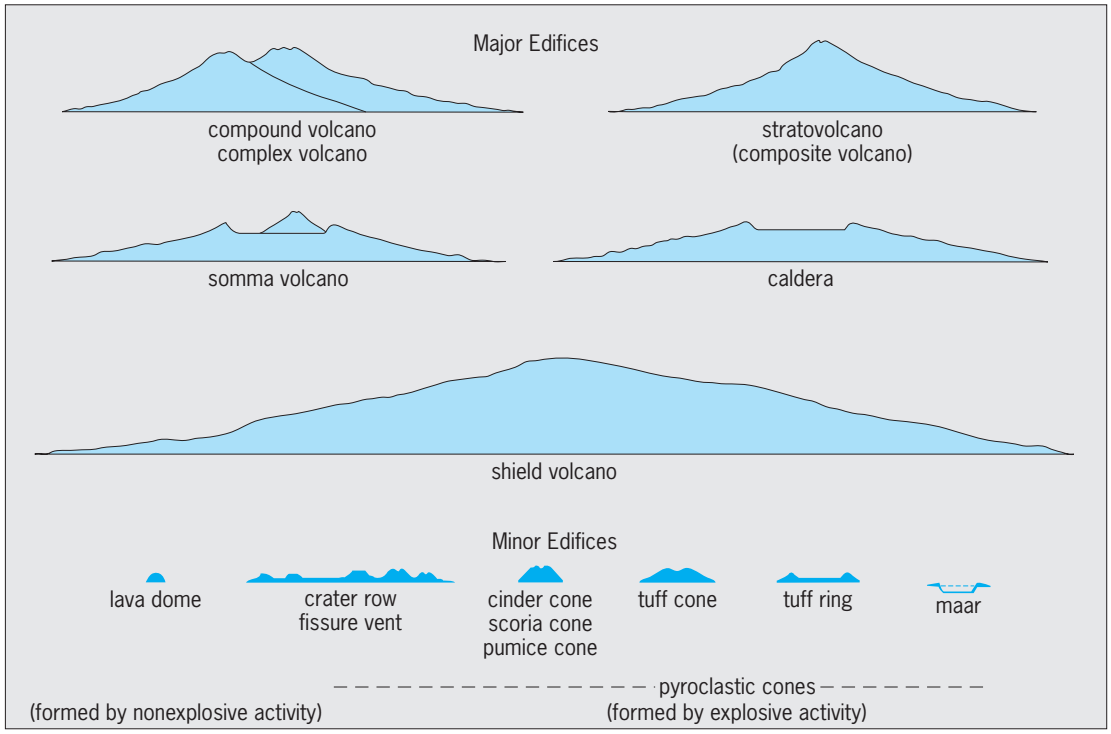
The only volcanic products other than basalt that are sufficiently fluid and voluminous to form extensive volcanic plateaus or plains are ash flows. Some of the larger examples of these ash-flow or ignimbrite plains, such as the central part of North Island of New Zealand, the Jemez Mountains region of New Mexico, and the Yellowstone Plateau of Montana-Wyoming, involve repeated eruptions of ash flows covering many thousands of square miles.

Minor structures. In general, minor volcanic structures originate or develop on or near the major edifice during a single eruptive event or a short span of activity, whereas the major edifices represent the result of repeated activity spanning many thousands or even millions of years.

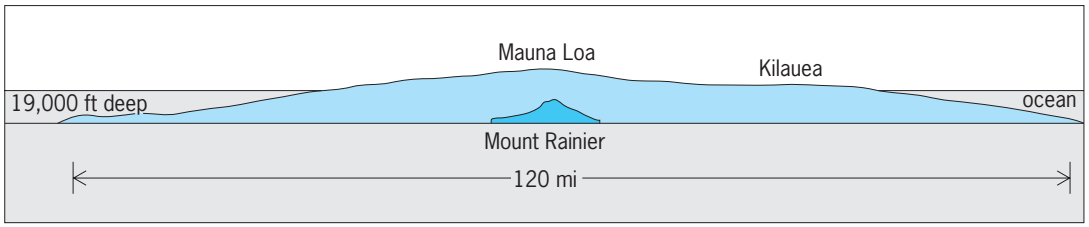
Lava too viscous to flow readily may accrete in or near the vent to form a steep-sided heap known as a lava or volcanic dome; such domes commonly develop following vigorous explosive activity at composite volcanoes (**Figs. 12** and **13**). Slender spires that thrust through apertures in such a dome are termed spines. The famous spine of Mont Pelée (Martinique, Lesser Antilles), formed during the eruption of 1902, reached a height of over 1280 ft (390 m), but like most such spines, was very short-lived. Other near-vent forms that may develop on the volcano include volcanic shield, cinder cone, spatter cone (pipe vent), spatter rampart (fissure vent), ash cone, tuff cone, and tuff ring (**Fig. 14**).

Major edifices are commonly modified by negative or depressional landforms. Small bowl-shaped depressions that are formed by explosion, or by failure of pyroclastic ejecta to accumulate directly above a vent, are known as craters. Most of them are found at the summit or on the flanks of volcanic cones, but some are well away from any cones.

Larger depressions at the summit of volcanoes are formed by collapse of the summit region as the support beneath it is removed by the rapid withdrawal of magma, usually by surface eruption but sometimes by subsurface migration of magma within the volcano. A depression formed by collapse is called a pit or collapse crater, or if larger than about 1.2 mi (2 km) in diameter, a caldera. Perhaps the best-known examples of calderas in the United States are Crater Lake, Oregon; Kilauea caldera, Hawaii; Yellowstone National Park (Wyoming); and Valles caldera, Jemez Mountains, New Mexico. Larger, though less regular and obvious, features of similar origin are known as



(a)



(b)

Fig. 10. Profiles of volcanoes. (*a*) Some common types of volcanoes; the relative sizes shown are only approximate and dimensions vary greatly within each group. For the major edifices, the vertical exaggeration is about 2:1; for the minor edifices, about 4:1 (*after T. Simkin and L. Siebert, Volcanoes of the World, 2d ed., Geoscience Press, 1994*). (*b*) Hawaiian shield volcanoes (Mauna Loa and Kilauea) compared with Mount Rainier (Washington), one of the larger composite volcanoes of the Cascade Range, drawn at the same scale (no vertical exaggeration) (*from R. I. Tilling et al., Eruptions of Hawaiian Volcanoes: Past, Present, and Future, USGS, 1987*).

volcanic-tectonic depressions. Like many calderas, their formation commonly, if not always, is associated with the explosive eruption of great volumes of ash flows. *See* IGNEOUS ROCKS; PETROLOGY.

Evolving landforms. Excluding postformation, gradual changes caused by weathering and erosion, volcanic landforms commonly undergo rapid changes during the course of an eruption, depending on its duration, type of eruptive activity, and so on. Such short-term, abrupt changes are dramatically illustrated by the changes in the summit configuration of Mount St. Helens (Washington) since 1980. Figure 13*a* is a photograph taken on May 17, 1980, from the north one day before the catastrophic eruption, showing the symmetrical form of a typical composite volcano (compare with Fig. 8). Figure 13*b* shows the same view, photographed after the catastrophic eruption of May 18, which removed the upper 1300 ft (400 m) of the summit and formed a 1.1 × 2.1 mi (1.7 × 3.3 km) amphitheater-shaped crater. Continued growth of the volcanic dome inside the crater (Fig. 12) could ultimately restore Mount St. Helens to its former shape. An intermediate stage in such a future restoration might be represented by Bezymianny volcano (Kamchatka, Russia), photographed in 1982 (Fig. 13*c*). The large crater, which formed in 1956 by a similar explosive eruption, is largely filled by lava of a growing volcanic dome.

Submarine volcanism. More than 80% of the Earth's crust is of volcanic origin, and about three-quarters of all the Earth's active volcanism takes place not on land but deep beneath the oceans. Such deep submarine volcanism occurs along the spreading ridges that zigzag for thousands of miles across the ocean floor, and it is exposed above sea level only in Iceland (**Fig. 15**). Because of the logistical difficulties in making direct observations posed by the great ocean depths, no deep submarine volcanic activity has

Fig. 11. A high cliff near Little Goose Dam, Washington, exposing a series of thick plateau basalts making up part of the Columbia River Plateau in the northwestern region of the United States. A car (circled) on the highway at the base of the cliff gives scale. (*D. A. Swanson, USGS*)

been actually observed during eruption by scientists. However, evidence that deep-sea eruptions are happening is clearly indicated by (1) eruption-induced earthquake activity recorded by seismic and acoustic monitoring networks; (2) the presence of high-temperature deep-ocean floor hydrothermal vents called smokers; (3) episodic short-lived but widespread hydrothermal discharges, measured and mapped as thermal and geochemical anomalies in the ocean water above the spreading ridges; and (4) the detection of new lava flows in certain segments of the oceanic ridge system (for example, Juan de Fuca and Gorda ridges off the coast of Oregon and Washington) where there had been none observed previously, as shown by repeated mapping of changes in the bathymetry and imaging of surface features. *See* HYDROTHERMAL VENT; MID-OCEANIC RIDGE.

Fig. 12. Ash and gas plume rises from the volcanic dome growing inside the large, amphitheaterlike crater formed by the May 1980 eruption of Mount St. Helens. In this picture (taken May 1983), the dome measured about 2800 ft (850 m) long, 2600 ft (800 m) wide, and 750 ft (230 m) high; the dome continued to grow until October 1986. (*L. Topinka, USGS*)

In contrast, shallow submarine volcanism has been directly observed and, in some cases, well studied. Volcanic eruptions in shallow water are very similar in character to those on land but, on average, are probably somewhat more explosive, owing to heating of water and resultant violent generation of supercritical steam. The glassy ash of cones formed in this way commonly is altered, probably by ordinary weathering processes, to brownish palagonite tuff. Such cones, like Diamond Head in Honolulu, usually have broader, flatter profiles than those characteristic of cinder cones.

Much of the ocean basin appears to be floored by basaltic lava, which, judging from its apparent density, is much less vesicular than the lavas of the basaltic cones that rise above it to form most of the oceanic islands. The pressure of the overlying water at great depths in the oceans may prevent the exsolution and explosive escape of gas from erupting lava and greatly reduce the vesiculation of the lava itself. Recent studies of submarine volcanism via deep submersible research vessels and dredged samples show that the volcanic landforms and flow features of basaltic lavas observed on land may be present on the ocean floor in addition to the ubiquitous pillow lavas. *See* OCEANIC ISLANDS.

Although pyroclastic materials in the ordinary sense probably seldom, if ever, form in deep water, lava flows may shatter on contact with the water, forming masses of glassy sand-size fragments resembling ash. Such material, called hyaloclastite, may form in water of any depth and is commonly associated with pillow lavas. Great volumes of hyaloclastite

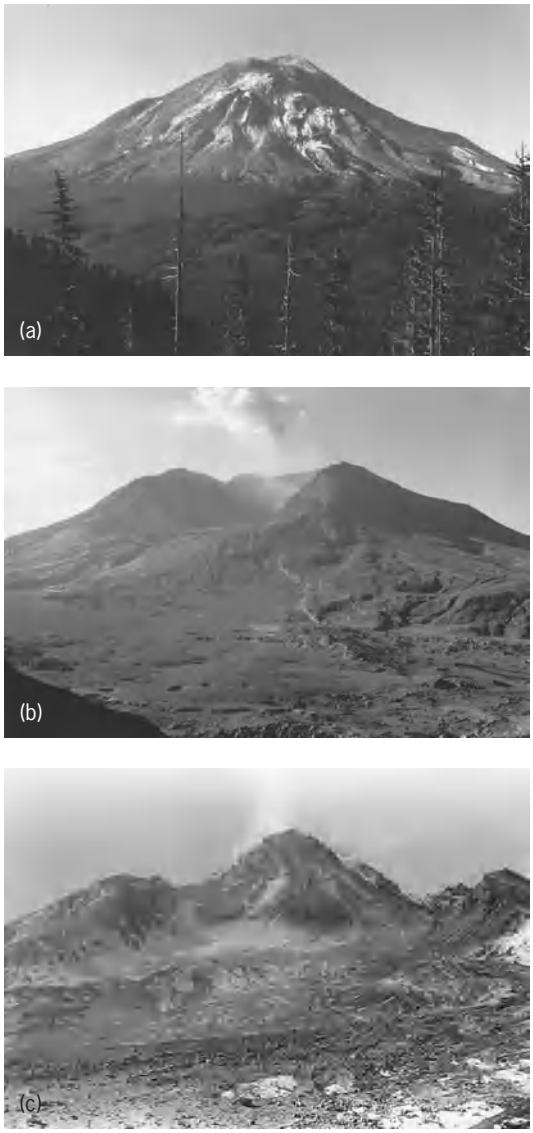


Fig. 13. Changes in the summit of Mount St. Helens. (*a*, *b*) Actual changes since 1980 (*photographs by Harry Glicken*, *USGS*). (*c*) A growing volcanic dome rises above the crater rim of Bezymianny volcano (Russia, 1982); this view may represent a possible future configuration of Mount St. Helens (*Institute of Volcanology*, *Petropavlovsk*, *Kamchatka*).

were formed in melt water by eruptions beneath glaciers in Iceland. Like ordinary basaltic ash, hyaloclastite readily alters to palagonite.

Fumaroles and hot springs. Vents at which volcanic gases issue without lava or after the eruption are known as fumaroles. They are found on active volcanoes during and between eruptions and on dormant volcanoes, persisting long after the volcano itself has become inactive. Fumarolic gases include water vapor, sulfur gases, hydrochloric and hydrofluoric acids, carbon dioxide and monoxide, and others in less abundance. They transport and deposit at the surface small amounts of many common metals. Temperatures of the escaping gases may reach 930–1470°F (500°–800°C), and the halogen gases and metals generally are found in the high-temperature fumaroles. Lower-temperature fumaroles, in which sulfur gases predominate along with steam, are called solfataras; still cooler ones liberating predominantly carbon gases are called mofettes. Researchers in submarine studies of the East Pacific Rise (20°N, off the mouth of the Gulf of California) and of the Gorda Ridge (~43°N, off the coast of Oregon) have photographed undersea fumaroles and sampled metallic sulfides they were emitting.

Fumaroles grade into hot springs and geysers. The water of most, if not all, hot springs is predominantly of meteoric origin, and is not water liberated from magma. Some hot springs appear to result simply from water circulating to warm regions at great depths in the Earth's crust, but in many the heat is of volcanic origin and the water may contain volcanic gases. Indeed, the heat may be derived wholly from rising hot volcanic gases. Some naturally heated water and steam are being developed commercially as geothermal energy to provide electricity or beneficial heat for space heating. *See* GEOTHERMAL POWER; GEYSER.

Distribution of volcanoes. Over 500 active volcanoes are known on the Earth, mostly along or near the boundaries of the dozen or so lithospheric plates that compose the Earth's solid surface (Fig. 15). These rigid plates, which range in thickness from 30 to 90 mi (50 to 150 km) and consist of both crustal and upper mantle material, form the lithosphere and move relative to one another above a hotter, more plastic zone in the mantle called the asthenosphere. *See* ASTHENOSPHERE; LITHOSPHERE; PLATE TECTONICS.

Lithospheric plates show three distinct types of boundaries: divergent or spreading margins—adjacent plates are pulling apart; convergent margins (subduction zones)—plates are moving toward each other and one is being destroyed; and transform margins—one plate is sliding horizontally past another. All these types of plate motion are well demonstrated in the Circum-Pacific region, in which many active volcanoes form the so-called Ring of Fire (Fig. 15).

Along spreading boundaries, new basaltic magma formed by partial melting of mantle material moves into the tensional fissures, solidifying as dikes or feeding lava flows on the ocean floor, accreting new material to the lithosphere. *See* RIFT VALLEY.

The Earth has remained more or less constant in size for the past few hundreds of millions of years; thus the accretion of the lithosphere by volcanism at the spreading boundaries must be compensated for by the destruction of lithosphere elsewhere. The compensatory consumption of lithosphere is assumed to take place along the convergent plate boundaries, where one plate plunges beneath (or is subducted by) another plate. Where the plate moves downward in a subduction zone, the lithosphere is depressed, forming deep troughs such as the Japan and Mariana trenches. As the sinking lithospheric slab enters the hot underlying mantle, it is partially melted, yielding magma that rises through the edge of the overlying plate to produce volcanic activity at

Fig. 14. Aerial view of a string of cinder cones ("crater row" example in Fig. 10*a*) along the fissure vent of the Lakagígar (Laki), Iceland, eruption of 1783, which resulted in the most voluminous outpouring of lava in recorded history. (*R. W. Decker, USGS*)

Fig. 15. Generalized diagram showing the major lithospheric plates and some of the active volcanoes of the Earth. Some scientists subdivide the Indo-Australian plate into the Indian and Australian plates.

the surface. The explosive volcanoes of the Circum-Pacific Ring of Fire, with their predominantly andesitic viscous magmas, have been generated by subduction processes. *See* PLATE TECTONICS.

Some volcanoes, however, are not associated with plate boundaries, and many of these so-called intraplate volcanoes form roughly linear chains in the interior parts of the oceanic plates, for example, the Hawaiian-Emperor, Austral, Society, and Line archipelagoes in the Pacific Basin. Intraplate volcanism also has resulted in voluminous outpourings of fluid lava to form extensive plateau basalts, or of more viscous and siliceous pyroclastic products to form ash flow plains.

Geologic data show that the Hawaiian-Emperor chain becomes progressively younger from northwest to southeast. This relationship, in conjunction with the parallelism in trend of the linear volcanic chain and plate motion, provides the best evidence for the so-called hot-spot or melting-spot model to account for the origin and development of the Hawaiian-Emperor and other linear volcanic chains. According to this model, as the northwest-moving Pacific plate passes over a hot, magma-generating spot in the mantle, the magma formed by partial melting rises through the oceanic lithosphere to form a volcano. This volcano is then gradually and steadily carried northwestward away from the hot spot, until it becomes isolated from the magma-generating source and dies, and a new volcano forms behind it over the hot spot. This process continues to form the entire linear volcanic chain. Although most scientists accept the geometric and dynamic aspects of the hot-spot model for the origin of linear intraplate volcanic chains, fundamental questions regarding the origin, persistence, location, or possible migration of hot spots remain. *See* EARTH; HOT SPOTS (GEOLOGY); MARINE GEOLOGY; PACIFIC ISLANDS.

Robert I. Tilling

Bibliography. M. H. Carr, *Volcanic Processes in the Solar System*, 1987; R. W. Decker and B. Decker, *Volcanoes*, 3d ed., 1998; R. W. Decker, T. L. Wright, and P. H. Stauffer (eds.), *Volcanism in Hawaii*, 1987; P. Francis, *Volcanoes: A Planetary Perspective*, 1993; B. L. Foxworthy and M. Hill, *Volcanic Eruptions of 1980 at Mount St. Helens: The First 100 Days*, 1982; S. L. Harris, *Fire Mountains of the West: The Cascades and Mono Lake Valcanoes*, 1988; W. J. Kious and R. I. Tilling, *This Dynamic Planet: The Story of Plate Tectonics*, 1996; P. W. Lipman and D. R. Mullineaux (eds.), *The 1980 Eruptions of Mount St. Helens, Washington*, 1981; G. A. Macdonald, *Volcanoes*, 1972; G. A. Macdonald, A. T. Abbott, and F. L. Peterson, *Volcanoes in the Sea: The Geology of Hawaii*, 2d ed., 1983; L. McClelland et al., *Global Volcanism, 1975–1985*, 1989; H. Sigurdsson et al. (eds.), *Encyclopedia of Volcanoes*, 2000; T. Simkin and L. Siebert, *Volcanoes of the World: A Regional Directory, Gazetteer, and Chronology of Volcanism During the Last 10,000 Years*, rev. ed., 1994; T. Simkin and R. S. Fiske, *Krakatau 1983: The Volcanic Eruption and Its Effects*, 1983; R. I. Tilling, *Eruptions of Mount St. Helens: Past, Present, and Future*, rev. ed., 1990; R. I. Tilling, *Volcanoes*, 1982; R. I. Tilling, C. C. Heliker, and T. L. Wright, *Eruptions of Hawaiian Volcanoes: Past, Present, and Future*, 1987; H. Williams and A. R. McBirney, *Volcanology*, 1979.

Volcanology

The scientific study of volcanic phenomena, especially the processes, products, and hazards associated with active or potentially active volcanoes. It focuses on eruptive activity that has occurred within the past 10,000 years of the Earth's history, particularly eruptions during recorded history. Strictly speaking, it emphasizes the surface eruption of magmas and related gases, and the structures, deposits, and other effects produced thereby. Broadly speaking, however, volcanology includes all studies germane to the generation, storage, and transport of magma, because the surface eruption of magma represents the culmination of diverse physicochemical processes at depth. This article considers the activity of erupting volcanoes and the nature of erupting lavas. For a discussion of the distribution of volcanoes and the surface structures and deposits produced by them *See* PLATE TECTONICS; VOLCANO.

Volcanoes and humans. From the dawn of civilization, volcanic eruptions have intruded into human affairs, producing death and destruction, bewilderment, fear, superstition, and, ultimately, scientific curiosity. Deities or supernatural events, directly or indirectly linked to volcanoes and eruptions, figure prominently in the legends and myths of civilizations that developed in or near regions of active volcanism. During the last 500 years, at least 200,000 people have lost their lives as a result of volcanic eruptions (**Table 1**).

Three eruptions in the 1980s appreciably increased public awareness of volcanic activity and of volcanology. The eruption of Mount St. Helens (Washington) on May 18, 1980 (**Fig. 1**), caused the worst volcanic disaster in the history of the United States, resulting in the loss of 57 lives. Yet, this eruption was much less destructive than other historic eruptions (Table 1). In March-April 1982, El Chichón, an obscure and largely forgotten volcano in southeastern Mexico, produced three major explosive bursts, which obliterated all settlements within a radius of about 5 mi (8 km) and perhaps caused more than 2000 deaths. Not only was this eruption the most destructive in Mexico's history, but some atmospheric scientists have claimed that the massive injection of sulfate aerosols into the stratosphere by El Chichón may have affected global climate, perhaps lowering the average temperature in the Northern Hemisphere by 0.18 or 0.36°F (0.1 or 0.2°C).

On November 13, 1985, a very small-volume eruption (0.007 mi^3 or 0.029 km^3) occurred at the summit crater of 17,680-ft-high (5389-m), glacier-capped Volcán Nevado del Ruiz, Colombia, the northernmost active volcano in the Andes. Despite the small amount of material erupted, the hot ejecta mixed

TABLE 1. Some historical volcanic eruptions

Volcano	Year	Estimated casualties	Principal causes of death
Merapi (Indonesia)	1006	>1,000	Explosions
Kelut (Indonesia)	1586	10,000	Lahars (mudflows)
Vesuvius (Italy)	1631	18,000	Lava flows, mudflows
Etna (Italy)	1669	10,000	Lava flows, explosions
Merapi (Indonesia)	1672	>300	Nuées ardentes, lahars
Awu (Indonesia)	1711	3,200	Lahars
Papandayan (Indonesia)	1772	2,957	Explosions
Laki (Iceland)	1783	10,000	Lava flows, volcanic gas, starvation*
Asama (Japan)	1783	1,151	Lava flows, lahars
Unzen (Japan)	1792	15,000	Lahars, tsunami†
Mayon (Philippines)	1814	1,200	Nuées ardentes, lava flows
Tambora (Indonesia)	1815	92,000	Starvation*
Galunggung (Indonesia)	1822	4,000	Lahars
Awu (Indonesia)	1856	2,800	Lahars
Krakatau (Indonesia)	1883	36,000	Tsunami†
Awu (Indonesia)	1892	1,500	Nuées ardentes, lahars
Mont Pelée, Martinique (West Indies)	1902	36,000	Nuées ardentes
Soufrière, St. Vincent (West Indies)	1902	1,565	Nuées ardentes
Taal (Philippines)	1911	1,332	Explosions
Kelut (Indonesia)	1919	5,000	Lahars
Lamington (Papua New Guinea)	1951	3,000	Nuées ardentes, explosions
Merapi (Indonesia)	1951	1,300	Lahars
Agung (Indonesia)	1963	3,800	Nuées ardentes, lahars
Taal (Philippines)	1965	350	Explosions
Mount St. Helens (United States)	1980	57	Lateral blast, mudflows
El Chichón (Mexico)	1982	>2,000	Explosions, nuées ardentes
Nevado del Ruiz (Colombia)	1985	>25,000	Mudflows
Unzen (Japan)	1991	41	Nuées ardentes
Pinatubo (Philippines)	1991	>300	Nuées ardentes, mudflows, ash fall (roof collapse)
Merapi (Indonesia)	1994	>41	Nuées ardentes from dome collapse
Soufriere Hills, Montserrat (West Indies)	1997	19	Nuées ardentes

*Deaths directly attributable to the destruction or reduction of food crops, livestock, agricultural lands, pasturage, and other disruptions of food chain.
†A Japanese word now commonly used worldwide for earthquake- or eruption-triggered seismic sea waves.

with melted snow and ice to generate highly mobile mudflows that swept down the steep drainages flanking the volcano. These mudflows killed more than 25,000 people downvalley, resulting in the second-worst volcanic disaster in the twentieth century (the worst is the 1902 Mont Pelée eruption at Martinique).

In 1991, two eruptions captured worldwide attention. On June 3, pyroclastic flows (nuées ardentes) triggered by the collapse of a new lava dome at the summit of Unzen Volcano, Kyushu, Japan, killed 41 people, including three volcanologists filming the volcanic activity. Intermittent weaker activity persisted at Unzen, and periodic collapses of the still-growing lava dome continued to pose a volcanic hazard to the city of Shimabara downslope from the volcano. During the course of the Unzen eruption, at least 12,000 people evacuated their homes temporarily. A much larger explosive eruption occurred at Mount Pinatubo, Luzon, Philippines, on June 15–16, 1991, after a dormancy of about 600 years. This eruption caused more than 300 fatalities and widespread destruction of structures, civil works, and cropland, and forced the evacuation of nearly 80,000 people, including 17,000 U.S. military personnel and their dependents stationed at Clark Air Base. Most of the fatalities attributed to the June 15 climactic eruption were caused by the collapse of roofs, laden with ash wetted by heavy rains of typhoon Yunya, which struck the island of Luzon at the same time. Since the 1991 eruption, destructive mudflows triggered by heavy rainfall during the monsoon seasons have caused additional fatalities and considerable property damage; such posteruption mudflows will pose a continuing volcano hazards into the twenty-first century, when the debris-choked valleys draining Mount Pinatubo are expected to reestablish the preeruption stream gradients. The Pinatubo eruption ranks as the second largest eruption in the world in the twentieth century, after that of Novarupta (Katmai), Alaska, in 1912. Moreover, Pinatubo injected into the stratosphere at least twice the volume of aerosols as did El Chichón in 1982, and the resultant stratospheric volcanic cloud affected the global climate until the mid-1990s. *See* CLIMATE HISTORY.

In the Caribbean region, Mont Pelée had been responsible for the world's worst volcanic disaster, in 1902 (Table 1). In mid-1995, Soufriere Hills, a volcano on the island of Montserrat (British West Indies), which had been dormant for more than three centuries, began to erupt. The most intense activity occurred in 1996–1997, mostly involving nuées ardentes triggered by a series of collapses of actively growing, unstable lava domes; sporadic weaker

activity has continued. This eruption is not large and has produced few fatalities, but it has caused tremendous socioeconomic and political impact. All of the island's means of livelihood and infrastructure have been lost, and the present population (3500–4000) is only about a third of that before the eruption. As of 2000, the British and Montserrat governments were still undecided about the long-range plans for the island's rehabilitation and possible return of the evacuated population.

On average, about 50 to 60 volcanoes worldwide are active each year. About half of these constitute continuing activity that began the previous year, and the remainder are new eruptions. Analysis of historic records indicates that eruptions comparable in size to that of Mount St. Helens or El Chichón tend to occur about once or twice per decade, and larger eruptions such as Pinatubo about once per one or two centuries. On a global basis, eruptions the size of that at Nevado del Ruiz in November 1985 are orders of magnitude more frequent.

Scientific inquiry. It was not until the nineteenth century that serious scientific inquiry into volcanic phenomena became part of the rapidly developing science of geology. Even though the building of a small observatory was completed in 1847 on the flank of Mount Vesuvius (Italy), modern volcanology perhaps began with the founding of well-instrumented observations at Asama Volcano (Japan) in 1911 and at Kilauea Volcano (Hawaii) in 1912. The Hawaiian Volcano Observatory, located on Kilauea's caldera rim, began to conduct systematic and continuous monitoring of seismic activity preceding, accompanying, and following eruptions, as well as other geological, geophysical, and geochemical observations and investigations. Operated by the U.S. Geological Survey (USGS), the Hawaiian Volcano Observatory pioneered and refined most of the commonly used volcano-monitoring techniques that are employed by other observatories studying active volcanoes elsewhere, principally in Iceland, Indonesia, Italy, Japan, New Zealand, Lesser Antilles (Caribbean), Philippines, and Kamchatka (Russia). In response to the Mount St. Helens eruption in 1980, the David A. Johnston Cascades Volcano Observatory was established by the U.S. Geological Survey. A sister observatory to the Hawaiian observatory, it monitors the eruptions of Mount St. Helens and serves as the center for the study of the other potentially active volcanoes of the Cascade Range (in California, Oregon, and Washington).

In March 1988, the USGS, in a cooperative program with the State of Alaska and the university of Alaska, established the Alaska Volcano Observatory, with facilities and staff in both Anchorage and Fairbanks.

In 1999 the USGS formally designated a long-term program of volcano-monitoring studies at Long Valley Caldera (east-central California) as the Long Valley Observatory. The last volcanic activity in the Long Valley region was about 200 years ago. Since May 1980 the caldera has exhibited measurable volcanic unrest, as seen by greatly increased seismicity activity and an accumulated ground uplift of nearly 2 ft (0.7 m). *See* CALDERA.

Fig. 1. Climactic eruption of Mount St. Helens on May 18, 1980, about 5 h after the beginning of activity. The plume of ash and gases reached an altitude of about 15 mi (24 km). (*Photograph by R. M. Krimmel, USGS*)

Nature of magmas. The eruptive characteristics, products, and resulting landforms of a volcano are determined predominantly by the composition and physical properties of the magmas involved in the volcanic processes (**Table 2**). Formed by partial melting of existing solid rock in the Earth's lower crust or upper mantle, the discrete blebs of magma consist of liquid rock (silicate melt) and dissolved gases. Driven by buoyancy, the magma blebs, which are lighter than the surrounding rock, coalesce as they rise toward the surface to form larger masses. *See* IGNEOUS ROCKS; LITHOSPHERE.

Concentration of volatiles. During its ascent, the magma enters zones of lower temperature and pressure and begins to crystallize, producing crystals suspended in the liquid—physically analogous to the formation of ice crystals when water begins to freeze. Other solid fragments may also be incorporated from

TABLE 2. Generalized relationships between magma composition, relative viscosity, and common eruptive characteristics

Magma composition	Relative viscosity	Common eruptive characteristics
Basaltic	Fluidal	Lava fountains, flows, and pools
Andesitic	Less fluidal	Lava flows, explosive ejecta, ashfalls, and pyroclastic flows
Dacitic-rhyolitic	Viscous	Explosive ejecta, ashfalls, pyroclastic flows, and lava domes

the walls and roof of the conduit through which the magma is rising. As crystallization progresses, volatiles and the more soluble silicate components are concentrated in the remaining liquid. *See* PHENOCRYST; XENOLITH.

At some point during magma ascent, decreasing confining pressure and increasing concentration of volatiles in the residual liquid initiate the separation of gas from the liquid. From that point on to its eruption, the magma consists of three phases: liquid, solid, and gas. Volcanic gases generally are predominantly water; other gases include various compounds of carbon, sulfur, hydrogen, chlorine, and fluorine. All volcanic gases also contain minor amounts of nitrogen, argon, and other inert gases, largely the result of atmospheric contamination at or near the surface.

In laboratory experiments, at a temperature of 2000°F (1100°C) and pressure of 5 kilobars (500 megapascals), a melt of rhyolitic composition can contain in solution about 10% (by weight) of water, a basaltic melt about 8%. At lower pressure, the solubility of water in any magma decreases correspondingly. With continued ascent, water and other volatiles in excess of their solubilities in the magma will exsolve, vesiculate, and ultimately increase "gas pressure" of the magma to provide the driving force for eruptions. This process may be compared with the uncorking of a bottle of champagne, especially if it has been shaken; the gas separates from the wine and forms bubbles, which in turn expand violently (explode) when the cork is removed suddenly.

The actual proportion of gas to lava liberated during eruptions cannot be directly determined; the amount of gas can be lower or higher than the values from laboratory experiments depending on the actual crystallization and degassing histories of the magmas. For many eruptions, volatiles measured in the lava constitute less than 1% (by weight) of the lava erupted during the same interval. In initially unsaturated magmas, high gas pressures may be developed by supersaturation of volatiles as their residual liquid phase becomes concentrated during crystallization. In recent decades, a combination of refined laboratory methods (to analyze melt inclusions and phenocrysts) and modern remote-sensing techniques (to measure volcanic gases in atmosphere) have been used to obtain data for well-studied eruptions. For example, data obtained at Mount St. Helens (1980), Redoubt Volcano, Alaska (1989), and Mount Pinatubo (1991) indicate that many magma systems were gas-saturated at the time of eruption. Some magmas, however, undergo considerable preeruption degassing during shallow subsurface storage and transport. *See* CRYSTALLIZATION; MAGMA.

Part of the gas liberated at volcanoes probably comes from the same deep-seated source as the silicate portion of the magma, but some may be of shallower origin. Part of the steam may be the result of near-surface oxidation of deep-seated hydrogen or interaction of hot magma or rock with ground water or geothermal fluids in the proximity of the reservoir-conduit system. Some of the oxidation of the sulfur gases must have taken place close to the surface. At some volcanoes, such as Vesuvius, the carbon gases may derive in part from reaction of the magma with limestone at shallow depth. Ammonia and hydrocarbon components present in some gases probably are derived from the organic constituents of sedimentary rocks near the surface.

In some eruptions, such as the 1924 eruption of Kilauea, Hawaii, temperatures are low and the gas is wholly or very largely steam. Fresh magmatic material may be entirely absent. In these phreatic (steam-blast) eruptions, the steam is simply heated ground water from the rocks adjacent to the magma reservoir and volcanic conduit. In other eruptions, such as that of Parícutin, Mexico, in 1943, the large volume of steam given off simultaneously with lava and smaller amounts of magmatic gas far exceeds the theoretical saturation limit of the magma, indicating that volatilized ground water was involved.

Physical properties. Temperatures of erupting magmas have been measured in lava flows and lakes, pyroclastic deposits, and volcanic vents by means of infrared sensors, optical pyrometers, and thermocouples. Reasonably good and consistent measurements have been obtained for basaltic magmas erupted from Kilauea and Mauna Loa volcanoes, Hawaii, and a few other volcanoes. Measured temperatures typically range between 2100 and 2200°F (1150 and 1200°C), and many measurements in cooling Hawaiian lava lakes indicate that the basalt becomes completely solid at about 1800°F (980°C). Perhaps the most reliable temperature determinations of Hawaiian lavas are obtained from experimentally calibrated geothermometers, involving the precise chemical analysis of the abundance of calcium or magnesium in the glass matrix (the quenched liquid phase). At Nyamlagira volcano, central Africa, in vents and in flows close to vents, temperatures ranged from about 1900 to 2000°F (1040 to 1095°C). Temperatures during the 1950–1951 eruption of Oshima, Japan, were in the same range. Locally, temperatures as high as 2550°F (1400°C) have been reported but not well documented; these anomalously high temperatures have been ascribed to the burning of volcanic gases in vents. *See* GEOLOGIC THERMOMETRY.

Temperature measurements on more silicic lavas are few and much less accurate because of the greater violence of the eruptions and the necessity of working at considerable distances as a safety precaution. In general, however, they suggest lower temperatures of eruption than those for mafic lavas. For example, for andesitic and more silicic lavas, available temperature estimates have ranged from about 1800 to 1330°F (980 to 720°C). Thermocouple measurements, made only 20 h after eruption of dacitic pyroclastic flows of Mount St. Helens (August 1980), yielded a temperature range of 1337–1540°F (725–838°C).

A few field measurements of the viscosity of flowing basic lavas have been made by means of penetrometers (instruments that measure the rate of

TABLE 3. General relationships between types of eruptions and some eruptive characteristics*

Volcanic explosivity index (VEI)	0†	1	2	3	4	5	6	7	8
Description	Nonexplosive	Small	Moderate	Mod-Large	Large	Very large			
Volume of ejecta, m^3‡	$<10^4$	10^4–10^6	10^6–10^7	10^7–10^8	10^8–10^9	10^9–10^{10}	10^{10}–10^{11}	10^{11}–10^{12}	$>10^{12}$
Ash plume height, km§	< 0.1	0.1–1	1–5	3–15	10–25	>25			
Qualitative description	Gentle, effusive			Explosive			Cataclysmic, paroxysmal, colossal		
Eruption type	Hawaiian	Strombolian		Vulcanian	Severe, violent, terrific Plinian		Ultraplinian		
Eruptions (total in file)	699	845	3477	869	278	84	39	4	0

*Modified from T. Simkin and L. Siebert, *Volcanoes of the World: A Regional Directory, Gazetteer, and Chronology of Volcanism During the Last 10,000 Years,* 2d ed., 1994.
†Nonexplosive "Hawaiian" eruptions are assigned a VEI of zero regardless of the volume of lava.
‡1 m^3 =35.3 ft^3.
§For VEIs 0–2, given as kilometers above crater; for VEIs 3–8, given as kilometers above sea level. 1 km = 0.6 mi.

penetration into liquid of a slender rod under a given strength of thrust) and by the shearing resistance to the turning of a vane immersed in the liquid. Viscosities also have been calculated from observed rates of flow in channels of known dimensions and slope or have been estimated from the chemical composition of the lava (by extrapolation of laboratory data on viscosities of simple molten silicate compounds). The best direct determinations of viscosity are vane-shear measurements in a lava lake at Kilauea, which yielded a best estimate of about 1900 poises (190 Pa · s) at a temperature of 2100°F (1150°C). Calculations based on rate of flow at both Kilauea and Mauna Loa, Hawaii, gave viscosities of 3000–4000 poises (300–400 Pa · s) for lava close to the vents, increasing at greater distances from the vents as the lava cools and stiffens to immobility. At Hekla, Iceland, a somewhat more silicic lava in the vent had a viscosity of about 10,000 poises (1000 Pa · s); and at Oshima, Japan, the lowest viscosities in two streams near the vent during the 1951 eruption were 5600 and 18,000 poises (560 and 1800 Pa · s), respectively. In general, empirical observations of flow behavior of more silicic lavas indicate that they are more viscous than basaltic lavas, but direct field measurements of their viscosities have not yet been made. *See* LAVA; MAGMA; PYROCLASTIC ROCKS; VISCOSITY.

Types of volcanic eruptions. The character of a volcanic eruption is determined largely by the viscosity of the liquid phase of the erupting magma and the abundance and condition of the gas it contains. Viscosity is in turn affected by such factors as the chemical composition and temperature of the liquid, the load of suspended solid crystals and xenoliths, the abundance of gas, and the degree of vesiculation. In very fluid lavas, small gas bubbles form gradually, and generally are able to rise through the liquid, coalescing to some extent to form larger bubbles, and escape freely at the surface with only minor disturbance. In more viscous lavas, the escape of gas is less free and produces minor explosions as the bubbles burst their way out of the liquid. In still more viscous lavas, at times there appears to be a tendency for the essentially simultaneous formation of large numbers of small bubbles throughout a large volume of liquid. The subsequent violent expansion of these bubbles during eruption shreds the frothy liquid into tiny fragments, generating explosive showers of volcanic ash and dust, accompanied by some larger blocks (volcanic "bombs"); or it may produce an outpouring of a fluidized slurry of gas, semisolid bits of magma froth, and entrained blocks to form high-velocity pyroclastic flows, surges, and glowing avalanches (nuées ardentes). Also, rising gases may accumulate beneath a solid or highly viscous plug, clogging the vent until it acquires enough pressure to cause rupture and attendant explosion that hurls out fragments of the disrupted plug.

Types of eruptions customarily are designated by the name of a volcano or volcanic area that is characterized by that sort of activity (**Table 3**), even though all volcanoes show different modes of eruptive activity on occasion and even at different times during a single eruption.

Eruptions of the most fluid lava, in which relatively small amounts of gas escape freely with little explosion, are designated Hawaiian eruptions. Most of the lava is extruded as successive, thin flows that travel many miles from their vents. Lava clots or spatter thrown into the air in fountains (**Fig. 2**) may remain fluid enough to flatten out on striking the ground, and commonly to weld themselves to form cones of spatter and cinder. An occasional feature of Hawaiian activity is the lava lake, a pool of liquid lava with convectional circulation that occupies a preexisting shallow depression or pit crater. Recent data, however, indicate that eruptions of Hawaiian volcanoes—the namesake for the "Hawaiian" (dominantly effusive) type of eruptive activity—may not be as nonexplosive as suggested by observations of the historical eruptions at the Kilauea and Mauna Loa volcanoes. Geological and dating studies of prehistoric volcanic ash deposits demonstrate that the frequency of explosive eruptions from Kilauea is comparable to those for many of the composite volcanoes of the Cascade Range of the Pacific Northwest.

Strombolian eruptions are somewhat more explosive eruptions of lava, with greater viscosity, and produce a larger proportion of pyroclastic material. Many of the volcanic bombs and lapilli assume rounded or drawn-out forms during flight, but commonly are sufficiently solid to retain these shapes on impact.

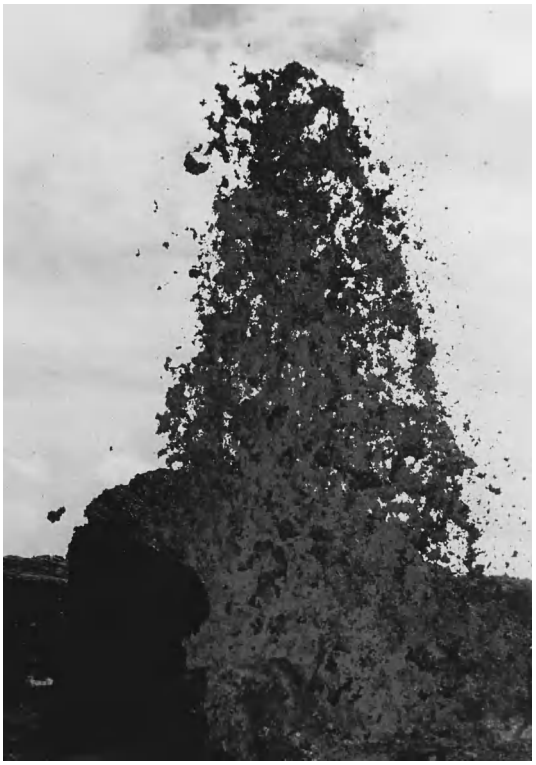


Fig. 2. Typical of "Hawaiian eruptions," an 80-ft-high (25-m) fountain of fluid basaltic lava plays during an eruption of Kilauea volcano, Hawaii, in 1973. Compare with the plinian eruption of Mount St. Helens in Fig. 1, involving much more viscous dacitic lava. (*Photograph by R. I. Tilling, USGS*)

Generally still more explosive are the vulcanian type of eruptions. Angular blocks of viscous or solid lava are hurled out, commonly accompanied by voluminous clouds of ash but with little or no lava flow.

Peléean eruptions are characterized by the heaping up of viscous lava over and around the vent to form a steep-sided hill or volcanic dome. Explosions, or collapses of portions of the dome, may result in glowing avalanches (nuées ardentes).

Plinian eruptions are paroxysmal eruptions of great violence—named after Pliny the Elder, who was killed in A.D. 79 while observing the eruption of Vesuvius—and are characterized by voluminous explosive ejections of pumice and by ash flows. The copious expulsion of viscous siliceous magma commonly is accompanied by collapse of the summit of the volcano, forming a caldera, or by collapse of the broader region, forming a volcano-tectonic depression. The term ultraplinian has been used occasionally by some volcanologists to describe especially vigorous plinian activity.

In contrast to the foregoing magmatic eruptions, some low-temperature but vigorous ultravulcanian explosions throw out fragments of preexisting volcanic or nonvolcanic rocks, accompanied by little or no new magmatic material. Certain explosion pipes and pits, known as diatremes and maars, have been produced by ultravulcanian explosions.

Volatilization of meteoric or ground water when it comes in contact with hot solid rocks in the vicinity of the volcano's magma reservoir causes the usually mild, but occasionally violent, disturbances known as phreatic explosions. No fresh magmatic material is erupted during phreatic (steam-blast) activity, which may precede major magmatic eruptions. The term "phreatomagmatic" describes a highly variable type of volcanic activity that results from the complex interaction between fresh magma/lava and subsurface or surface water (ground water, hydrothermal water, meteoric water, seawater, lake water).

The differences between the different types of eruptions are gradational but ultimately are dependent on the variation in explosivity and "size" of the eruption. Some volcanologists have proposed the volcanic explosivity index (VEI) to attempt to standardize the assignment of the size of an explosive eruption, using volume of eruptive products, duration, height of ash plume, and other criteria. Table 3 shows the general relationships and the necessarily arbitrary distinctions between eruption type, explosivity, and eruptive volume for nearly 6300 eruptions during the Holocene (the past 10,000 years of the Earth's history), for which such information is known or can be reasonably estimated. Of these eruptions, about 11% are nonexplosive Hawaiian-type eruptions (always assigned a VEI of 0 regardless of eruptive volume). Explosive eruptions considered to be small to moderate in size (VEI 1 to 2) constitute about 69% of all eruptions and are gradationally classified as Hawaiian, strombolian, or vulcanian. Only 127 plinian or ultraplinian eruptions rate VEIs of 5 or greater (very large). The May 18, 1980, eruption of Mount St. Helens rated a VEI of 5, but just barely. During the past 10,000 years, only four eruptions rated VEIs of 7, including the 1815 eruption of Tambora (Indonesia), the largest known eruption in the world in recorded history. Several of the large pre-Holocene caldera-forming eruptions would qualify for VEI ratings of 8. The VEI scale, like the Richter magnitude scale for earthquakes, is open-ended; that is, there is no maximum rating. However, to date, no eruptions of VEI 9 size (eruptive volume of 10^4 km^3) have been recognized, and magma reservoirs of such volume are virtually unknown in the geologic record. *See* HOLOCENE.

Monitoring active volcanoes. A major component of the science of volcanology is the systematic and, preferably, continuous monitoring of active and potentially active volcanoes. Scientific observations and measurements—of the visible and invisible changes in a volcano and its surroundings—between eruptions are as important, perhaps even more crucial, than during eruptions. Measurable phenomena important in volcano monitoring include earthquakes; ground movements; variations in gas compositions; and deviations in local gravity, electrical, and magnetic fields. These phenomena reflect pressure and stresses induced by subsurface magma movements and or pressurization of the hydrothermal envelope surrounding the magma reservoir.

Seismicity and ground deformation. The monitoring of volcanic seismicity and ground deformations before, during, and following eruptions has provided

the most useful and reliable information. From the many decades of study of the active Hawaiian volcanoes and volcanoes elsewhere, it has been determined that a volcano generally undergoes measurable ground deformation when magma is fed into its near-surface reservoir-conduit system.

Vertical and horizontal ground displacements and slope changes can easily be detected and measured precisely by existing geodetic techniques. Slope changes can be measured with a precision of a microradian or less by various electronic-mechanical "tilt-meters," or with somewhat less precision by leveling of short-sided arrays of benchmarks. Vertical displacements of benchmarks on a volcano, relative to a reference benchmark (or tide gage) unaffected by the volcano, can be determined to a few parts per million by leveling surveys. Horizontal distance changes between benchmarks can be measured with similar precision by various electronic distance measurement instruments.

Advances in geodetic applications of satellite positioning and other forms of space geodesy, especially the Global Positioning System (GPS), suggest that conventional ground-deformation monitoring techniques will be supplanted by satellite-based monitoring systems, if the acquisition and maintenance costs can be decreased substantially. Toward the end of the twentieth century, significant progress was made in the development and testing of near-real-time GPS monitoring networks at selected volcanic systems [for example, Augustine Volcano (Alaska), Long Valley Caldera (California)]. Another promising satellite-based geodetic technique is the InSAR (interferometric synthetic aperture radar) method, which is capable of detecting ground movements over extensive areas (that is, over the entire volcano). This technique involves the interferometric analyses of one or more pairs of SAR images acquired at time intervals; coherent ground movement is revealed by ringlike or other regular anomalies reflecting differences (interferences) in topography between the time-separated pair of satellite images being compared. The InSAR technique, while not yet a routine volcano-monitoring tool, has been shown to be successful at Yellowstone (Wyoming) and Long Valley calderas and some other volcanoes [such as, Mount Etna (Italy) and Okmok (Alaska)]. *See* GEODESY; SATELLITE NAVIGATION SYSTEMS; SEISMOGRAPHIC INSTRUMENTATION.

The ground deformation is related to, and accompanied by, intense earthquake activity reflecting the subsurface ruptures of the confining rocks of the expanding volcanic reservoir in response to the increased pressure exerted by infilling magma. Modern volcano observations employ well-designed seismic networks to monitor volcanic seismicity continuously in order to track subsurface movement of magma between and during eruptions. For well-studied volcanoes, experience has shown that premonitory seismicity usually provides the earliest signals of impending activity. Great advances have been made in volcano seismology since the early 1980s; in particular, recent studies demonstrate that the occurrence of, and variations in, long-period (low-frequency) events and volcanic tremor provide valuable insights in inferring the movement of magma and/or hydrothermal fluids before and during eruptions.

The number of earthquakes and the magnitude of ground deformation gradually increase as the magma reservoir swells or inflates until some critical strength threshold is exceeded and major, rapid migration of magma ensues to feed a surface eruption or a subsurface intrusion (magma drains from the reservoir and is injected into another part of the volcanic edifice without breaching the surface). With the onset of eruptive or intrusive activity, pressure on the "volcanic plumbing system" is relieved and the reservoir abruptly shrinks or deflates, causing flattening of slope (tilt), reduction in vertical or horizontal distances between surface points, and decrease in earthquake frequency. One or more of these inflation-deflation cycles may take place during an eruption, depending on its duration (**Figs. 3** and **4**). *See* EARTHQUAKE; SEISMOLOGY.

Gas emission and other indicators. As magma ascends, the emission rate or composition of gases exsolving from it and those generated by interaction of ground water and hot rock or magma may change with time. The systematic measurement at volcanic vents and fumaroles of such variations, though still largely experimental, shows much promise as another volcano monitoring tool. With modern methods (such as gas chromatography and mass spectroscopy), the composition of gas can be determined routinely soon after its collection. A field gas chromatograph has been developed and has been tested at several active volcanoes, including Etna (Italy), Kilauea (Hawaii), and Merapi (Indonesia). At Kilauea, volcanic gases have been monitored at 25 sites by sampling twice a week and analyzing chromatographically more than 10 different gas species. *See* GAS CHROMATOGRAPHY.

Some success has been obtained at Etna, Mount St. Helens, Galunggung (Indonesia), Mount Pinatubo (Luzon, Philippines) and other volcanoes in measuring the fluctuation in the emission rate of sulfur dioxide by means of a correlation spectrometer (COSPEC) in both ground-based and airborne modes. The emission rate of carbon dioxide (CO_2) was monitored remotely, using a modified infrared spectrophometer during some of the 1980–1981 eruptions of Mount St. Helens. Another recently developed remote gas-monitoring technique uses Fourier transform infrared (FTIR) spectroscopy for measurement of sulfur dioxide (SO_2), the ratio of sulfur dioxide to hydrochloric acid (SO_2/HCl), and silicon tetrafluoride (SiF_4) in volcanic plumes; this technique has been applied at the Kilauea (Hawaii), Etna and Vulcano (Italy), and Asama and Unzen (Japan) volcanoes. The continuous monitoring of emission at selected sites, together with periodic regional surveys, has demonstrated that the rate of carbon dioxide emission in the Long Valley Caldera region greatly increased in 1989. In areas of highest emission, the high CO_2 concentrations in the soil are killing trees by denying their roots oxygen and by interfering with

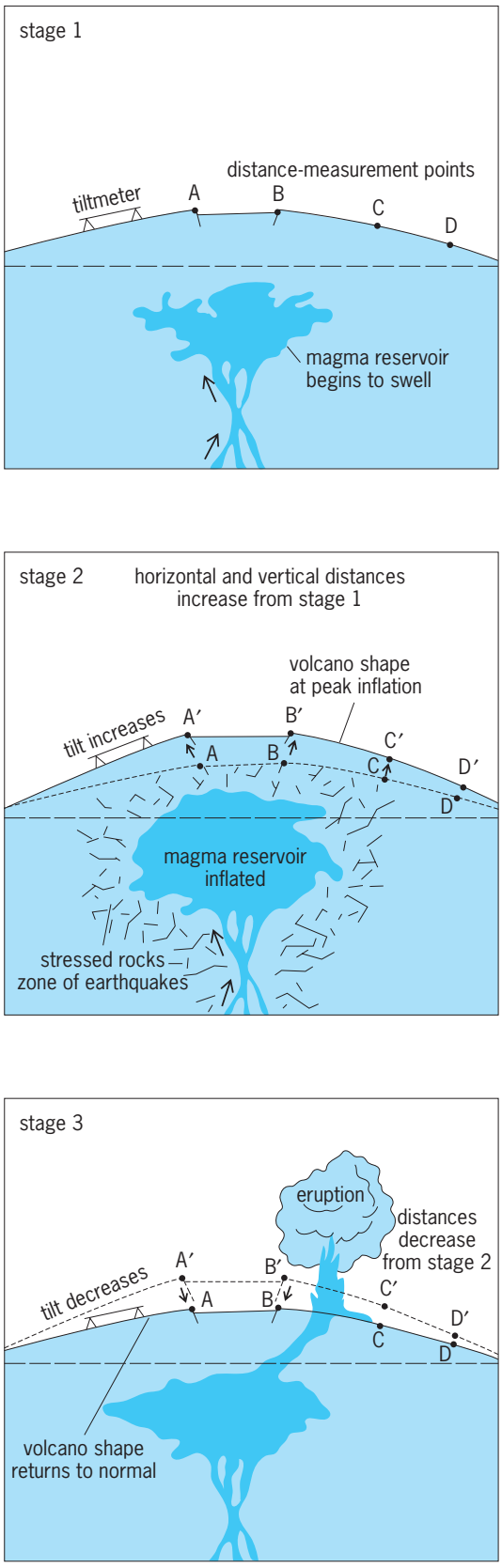


Fig. 3. Schematic diagrams showing three commonly observed stages in the course of an inflation-deflation cycle during a typical Hawaiian eruption. At stage 1, inflation begins; it peaks at stage 2. Stage 3 is the eruption-deflation stage. (*After R. I. Tilling, Monitoring Active Volcanoes, USGS, 1983*)

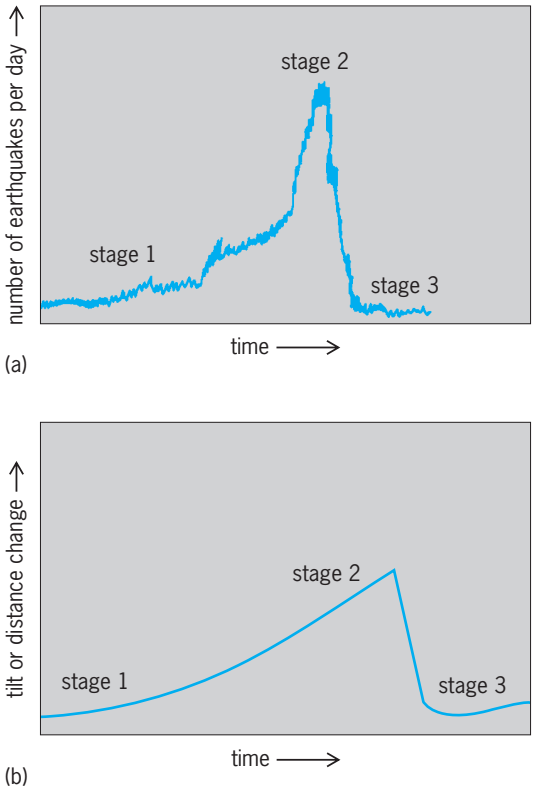


Fig. 4. Idealized graphs of (*a*) earthquake frequency and (*b*) tilt or distance changes as a function of time during the three stages of Fig. 3. (*After R. I. Tilling, Monitoring Active Volcanoes, USGS, 1983*)

nutrient uptake. Large but gradually declining volumes of CO_2 gas continue to seep from this volcanic system. The periodic determination of the absorption of certain gases (such as chlorine and sulfur) in alkaline solutions provides a time-integrated measure of their emission.

A significant advance in the monitoring of volcanic gas emission was the accidental discovery in the early 1980s that, with modification in computer algorithm and data processing, the Total Ozone Mapping Spectrometer (TOMS) instrument aboard the *Nimbus 7* satellite can be used to measure the output of sulfur dioxide during an eruption and to track the movement and, ultimately, the dissipation of the sulfur dioxide–containing stratospheric volcanic clouds. The combination of COSPEC and TOMS measurements of sulfur dioxide provides a powerful volcano-monitoring tool. *See* SPECTROSCOPY.

Periodic measurement of gas composition and emission rate, though providing important information about the volcanic system, does not, however, give the critical data on short-term or continuous fluctuations—as can seismic and ground deformation monitoring—that might augur an impending eruption. Some attempts have begun to develop continuous real-time monitoring of certain of the relatively abundant and nonreactive gases, such as hydrogen, helium, and carbon dioxide. Particularly encouraging is the continuous monitoring of hydrogen emission, utilizing electrochemical sensors and

satellite telemetry, that has been tested at Mount St. Helens, Kilauea, Mauna Loa, and Long Valley caldera (California) in the United States, and at Vulcano and other sites in Italy. Nonetheless, gas-monitoring techniques must be considered largely experimental, in large measure because of the fugitive nature of gases, their complex pathways from magmatic source to measurement site, and interactions with the hydrology of the volcano and its surroundings.

Even more experimental than gas monitoring are techniques involving measurements of changes in the thermal regime, in gravitational or geomagnetic field strength, and in various measurable "geoelectrical" parameters, such as self-potential, resistivity, and very low-frequency signals from distant sources. All of these methods are premised on the fundamental model that an inflating (or expanding) magma reservoir causes thermal anomalies and related magma-induced changes in bulk density, electrical and electromagnetic properties, piezoelectric response, and so forth, of the volcanic edifice.

Primarily because of the complicating effects of near-surface convective heat transfer, thermal monitoring of volcanoes generally has proved to be nondiagnostic, although monitoring of temperature variations in crater lakes at some volcanoes—Taal (Philippines) and Kelut (Indonesia)—has given useful information.

Although modern gravity meters can detect changes as small as 5 microGal (0.05 micrometer/s^2), interpretation of the results of gravity surveys is highly dependent on the ability to discriminate between elevation-induced and mass-difference changes. Preliminary analysis of some electrical self-potential data at Kilauea suggests that this technique may provide more advanced notice of impending eruptions than seismic and ground deformation monitoring. However, much more data and testing are required before geoelectrical monitoring techniques can be considered to be as routinely and universally applicable as seismic and ground deformation methods. *See* EARTH, GRAVITY FIELD OF; GEOELECTRICITY; GEOPHYSICAL EXPLORATION.

Volcanic hazards mitigation. Volcanoes are in effect windows into the Earth's interior; thus research in volcanology, in contributing to an improved understanding of volcanic phenomena, provides special insights into the chemical and physical processes operative at depth. However, volcanology also serves an immediate, perhaps more important, role in the mitigation of volcanic and related hydrologic hazards (mudflows, floods, and so on). Progress toward hazards mitigation can best be advanced by a combined approach. One aspect is the preparation of comprehensive volcanic hazards assessments of all active and potentially active volcanoes, including a volcanic risk map for use by government officials in regional and local land-use planning to avoid high-density development in high-risk areas. The other component involves improvement of predictive capability by upgrading volcano-monitoring methods and facilities to adequately study more of the most dangerous volcanoes. An improved capability for eruption forecasts

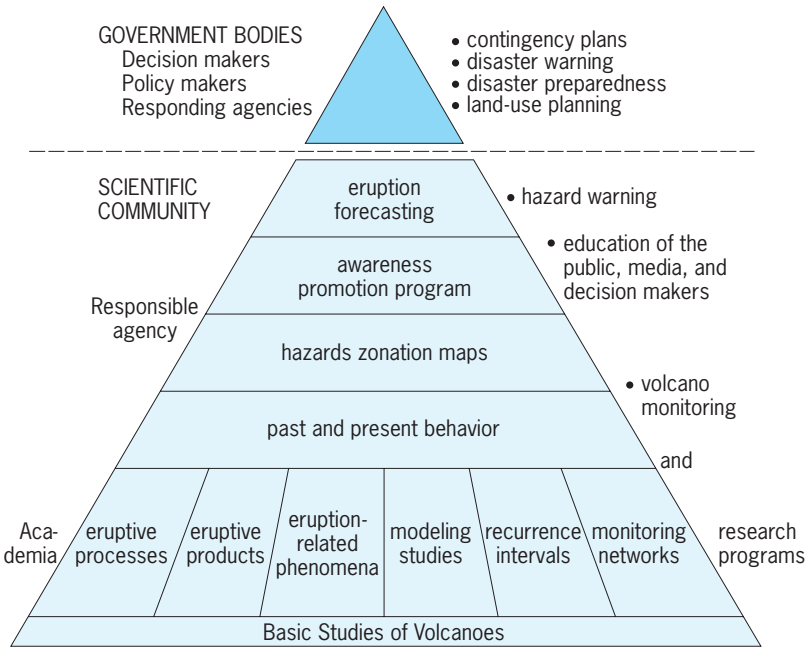


Fig. 5. Schematic diagram showing the general process of volcanic hazards assessment and development of mitigation strategies. (*After R. I. Tilling, ed., Volcanic Hazards, American Geophysical Union, 1989*)

and predictions would permit timely warnings of impending activity, and give emergency-response officials more lead time for preparation of contingency plans and orderly evacuation, if necessary.

Both these approaches must be buttressed by long-term basic field and laboratory studies to obtain the most complete understanding of the volcano's prehistoric eruptive record (recurrence interval of eruptions, nature and extent of eruptive products, and so on). Progress in mitigation of volcanic hazards must be built on a strong foundation of basic and specialized studies of volcanoes (**Fig. 5**). The separation of the apex from the rest of the triangle reflects the fact that "decision makers" must consider administrative and socioeconomic factors in addition to scientific information from volcanology. Nonetheless, volcanologists and other physical scientists—individually and collectively—must step out of their traditional academic roles and work actively with social scientists, emergency-management officials, educators, the news media, and the general public to increase public awareness of volcanoes and their potential hazards. Recent volcanic crises and disasters have shown that scientific data—no matter how much or how precise—serve no purpose in volcanic-risk management unless they are communicated effectively to, and acted upon in a timely manner by, the civil authorities.

One of the best examples of good volcanic hazards assessment and eruption forecasts is provided by the Mount St. Helens experience. As early as 1975, scientists predicted that this volcano was the one in the conterminous United States most likely to reawaken and to erupt, possibly before the end of the century. This prophecy was followed in 1978 by a detailed analysis of the types, magnitudes, and areal extents

Fig. 6. Climatic eruption of Mount Pinatubo (Luzon, Philippines) on June 15, 1991, with U.S. Clark Air Base in the foreground. The base of the eruption column measures about 16 mi (25 km) across. (*USAF photograph by Robert Lapointe*)

of potential volcanic hazards that might be expected from a future eruption of Mount St. Helens. The volcano erupted catastrophically on May 18, 1980, and the volcanic events and associated hazards largely followed the scenario outlined earlier. Because of improved volcano monitoring with the establishment of the David A. Johnston Cascades Volcano Observatory, all of the eruptions of Mount St. Helens since June 1980 have been predicted successfully.

The response by United States and Philippine volcanologists to the reawakening of Mount Pinatubo in early April 1991 (**Fig. 6**) constitutes another successful case in volcanic hazards mitigation. Seismic and volcanic gas monitoring studies prompted scientists to recommend to the U.S. and Philippine governments evacuation of the surrounding region and so more than 200,000 people were moved. About 36 h later, the cataclysmic eruption of June 15–16 took place, obliterating the region within a 5–6-mi (8–10-km) radius. Given the huge volume of the eruption (preliminary estimate is about 0.5 mi^3 or 2 km^3) and the widespread devastation, the death toll almost certainly would have been tens of thousands, rather than hundreds.

In contrast, the other two major volcanic disasters in the 1980s—El Chichón (1982) and Nevado del Ruiz (1985) [Table 1]—are sobering examples of failures in mitigation of hazards. In the case of El Chichón, the eruption came as a surprise because there were no geologic data about its eruptive history and no preeruption volcanic hazards assessment or monitoring. The Ruiz catastrophe, however, is a more tragic case because considerable geologic data existed; warning signs were recognized a year earlier and limited volcano monitoring was initiated; and a preliminary hazards-zonation map, produced more than a month before the destructive eruption, correctly pinpointed the areas of greatest hazard. Sufficient warnings were given by the scientists on site at Ruiz, but for reasons still not clear, effective evacuation and other emergency measures were not implemented by the government authorities.

In contrast, the volcanic crisis at Rabaul Caldera (Papua New Guinea) in the mid-1980s offers an excellent example of how an effective hazards-mitigation program can save lives, even though the anticipated eruption came a decade later than the original forecast. Beginning in the early 1970s, Rabaul began to exhibit signs of volcanic unrest, as seen by periodic earthquake swarms beneath the caldera and by ground uplift. During 1983 and 1984, the volcano's activity increased dramatically, with the monthly counts of earthquakes approaching 10,000 late in 1983 (**Fig. 7***a*). This sudden escalation in seismicity was accompanied by sharply increasing rates of ground deformation, as indicated by geodetic monitoring. In October 1983, after considering socioeconomic factors and the scientific information provided by the Rabaul Volcano Observatory on the status of the volcanic unrest, Rabaul government officials declared a stage-2 alert, which implied that an eruption would occur within a few months. In response to this declaration, the citizens of Rabaul Town were made aware of the need to prepare for possible evacuation, staging areas for evacuation were designated, certain roads were widened to serve as evacuation routes, and several evacuation drills were conducted.

The rate of seismicity and ground deformation continued to increase for another 6 months following the declaration of the stage-2 alert, but then the level of unrest declined rapidly (Fig. 7*a*). The expected eruption did not happen, and the officials ended the alert in November 1984. For the next 10 years, caldera activity fluctuated with relatively low levels, but slightly higher than pre-1983 rates. Then, on September 19, 1994, following only 27 hours of precursory seismic and ground-deformation activity, explosive eruptions began at Vulcan and Tavurvur, the two vents on opposite sides of the caldera that have been the sites of previous historical eruptions (Fig. 7*b*). While Rabaul Town suffered massive destruction (principally from ash accumulated and roof collapse) and over 50,000 people were displaced, fewer than 10 people were killed (several from automobile accidents)—quite remarkable considering the rapid onset of the eruption with virtually no public warning to the population. The main reason for the low fatalities of the 1994 eruption was that people quickly "self-evacuated" with the first light ash falls. The people apparently had learned their lessons well years earlier, from the heightened awareness of volcanic hazards generated by the 1983–1985 volcanic crisis as well as from memories or stories of the 1937 eruption (similarly involving eruptive outbreaks at Vulcan and Tavurvur) among the long-time residents and their families.

Volcanic ash and aviation safety. During recent decades, with the advent of high-performance jet engines, an unrecognized volcano hazard emerged: in-flight encounters between jet aircraft and volcanic ash clouds produced by powerful explosive eruptions. This hazard stems from the following: (1) volcanic ash clouds are not detectable by the aircraft's onboard radar instrumentation; and (2) if the gritty,

jagged ash particles are ingested into the aircraft's jet engines, the high operating temperatures can partially melt the ash. Severe abrasion and ash accumulation within the engine, along with adherence of melted ash to critical engine parts and openings, combine to degrade engine performance and, at worst, can cause engine flameout and power loss. Since the early 1970s, more than 60 volcanic ash–aircraft encounters have occurred, with several of the aircraft experiencing total power loss and requiring emergency landings; fortunately, to date, no fatal crashes have resulted from such encounters. However, many millions of dollars of damage to aircraft have been incurred; for example, an encounter between a Boeing-747 jetliner and the ash cloud from an eruption of Redoubt Volcano (Alaska) in December 1989 required more than $80 million dollars to replace all four engines and repair other damage. Volcanologists worldwide are now working closely with the air-traffic controllers, civil aviation organizations, and the air-carrier industry to mitigate the hazards of volcanic ash to civil aviation. A part of this effort involves the operation of nine regional Volcanic Ash Advisory Centers around the world to provide early warning of explosive eruptions and to issue advisories of potentially dangerous volcanic ash clouds produced by them.

Impact of explosive volcanism on global climate. Large explosive eruptions that eject copious amounts of volcanic aerosols into the stratosphere also can affect climate on a global basis. For example, long-lingering stratospheric volcanic clouds—for example, from the great 1883 eruption of Krakatau (between Java and Sumatra, Indonesia), the 1963 eruption of Agung volcano (Bali, Indonesia), and the 1982 eruption of El Chichón (Chiapas, Mexico)—produced spectacular sunrises and sunsets all over the Earth for many months because of the interaction of suspended aerosols and the atmosphere. Studies indicate that the sulfate aerosols in volcanic clouds form a layer of sulfuric acid droplets. This layer tends to cool the troposphere by reflecting solar radiation, and to warm the stratosphere by absorbing radiated Earth heat; in general, the combined effect is to lower the Earth's surface temperature.

The 1980 Mount St. Helens eruption apparently had minimal climatic impact, producing at most only a decrease of 0.18°F (0.1°C) in average temperature for the Northern Hemisphere. In contrast, the 1982 El Chichón and the 1991 Pinatubo eruptions lowered temperature by 0.36–0.9°F (0.2–0.5°C). Significantly, the Mount St. Helens magma contains much less sulfur [<0.01 wt % of sulfur trioxide (SO_3)] than those of El Chichón or Pinatubo, 1–4 and 0.3–0.5 wt % respectively. Perhaps the best-known example of eruption-induced climate change is that associated with the eruption of Tambora volcano (Indonesia) in 1815, the most powerful eruption in historical time. Average temperature in the Northern Hemisphere was lowered by as much as 0.9°F (0.5°C), and parts of the United States and Canada experienced unusual summer frosts and loss of crops; 1816 was known as the year without summer. On a local scale, the fallout from gas-rich eruption plumes and associated acid rains can cause damage to crops and exposed metal parts of dwellings. *See* ACID RAIN; AEROSOL; AIR POLLUTION.

Robert I. Tilling

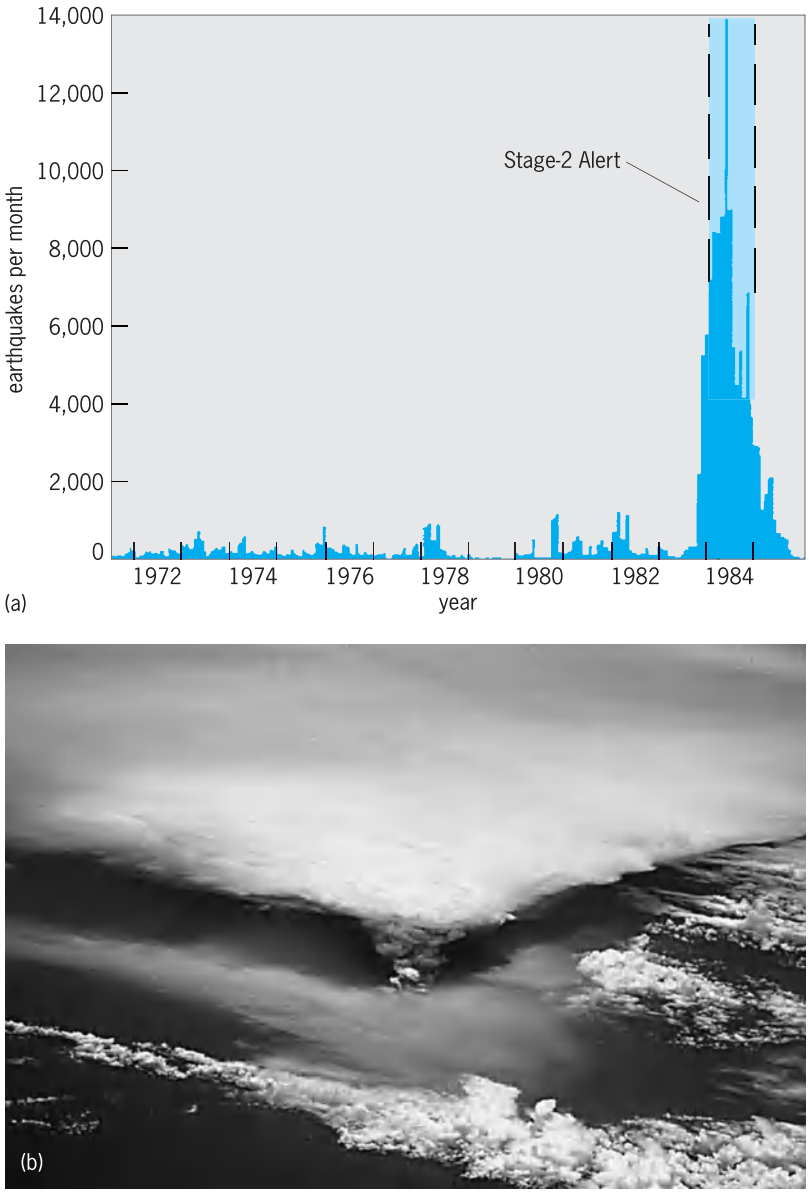


Fig. 7. Activity at Rabaul Caldera, Papua New Guinea. (*a*) Earthquakes per month from July 1971 through late 1985. The sharp increase in earthquakes in late 1983, combined with a cumulative uplift of about 2 m (6.6 ft) since 1973, prompted government officials to declare a stage-2 alert (in effect October 29, 1983, to November 22, 1984), in anticipation of a possible eruption within a few months. However, the eruption did not occur until mid-September 1994, about a decade after the alert was lifted. (*After R. I. Tilling, The role of monitoring in forecasting volcanic events, in B. McGuire et al., eds., Monitoring Active Volcanoes: Strategies, Procedures and Techniques, pp. 369–401, UCL Press Limited, London, 1995*) (*b*) Photograph, taken by the space shuttle astronauts, of the 18-km-high (11-mi) volcanic plume produced by the Rabaul eruption, about 24 hours after it began on September 19, 1994. (*NASA*)

Bibliography. R. J. Blong, *Volcanic Hazards: A Sourcebook on the Effects of Eruptions*, 1984; F. R. Boyd, Jr. (ed.), *Explosive Volcanism: Inception, Evolution, and Hazards*, 1984; T. J. Casadevall (ed.), *Volcanic Ash and Aviation Safety: Proceedings of the 1st International Symposium on Volcanic Ash and Aviation Safety*, 1994; R. W. Decker and

B. Decker, *Volcanoes*, 3d ed., 1998; R. W. Decker, T. L. Wright, and P. H. Stauffer (eds.), *Volcanism in Hawaii*, 1987; B. L. Foxworthy and M. Hill, *Volcanic Eruptions of 1980 at Mount St. Helens: The First 100 Days*, 1982; P. Francis, *Volcanoes: A Planetary Perspective*, 1993; R. W. Johnson, *Volcanic Eruptions and Climate Change*, 1993; P. W. Lipman and D. R. Mullineaux (eds.), *The 1980 Eruptions of Mount St. Helens, Washington*, 1981; G. A. Macdonald, *Volcanoes*, 1972; W. J. McGuire, C. R. J. Kilburn, and J. B. Murray (eds.), *Monitoring Active Volcanoes: Strategies, Procedures, and Techniques*, 1995; C. G. Newhall and R. S. Punongbayan (eds.), *Fire and Mud: Eruptions and Lahars of Mount Pinatubo, Philippines*, 1996; R. Scarpa and R. I. Tilling, *Monitoring and Mitigation of Volcano Hazards*, 1996; P. D. Sheets and D. K. Grayson (eds.), *Volcanic Activity and Human Ecology*, 1979; H. Sigurdsson et al. (eds.), *Encyclopedia of Volcanoes*, 2000; T. Simkin and L. Siebert, *Volcanoes of the World: A Regional Directory, Gazetteer, and Chronology of Volcanism During the Last 10,000 Years*, rev. ed., 1994; R. I. Tilling, *Eruptions of Mount St. Helens: Past, Present, and Future*, rev. ed., 1990; R. I. Tilling (ed.), *Volcanic Hazards*, 1989; R. I. Tilling, C. C. Heliker, and T. L. Wright, *Eruptions of Hawaiian Volcanoes: Past, Present, and Future*, 1987; R. I. Tilling and P. W. Lipman, *Lessons in Reducing Volcano Risk*, 1993; H. Williams and A. R. McBirney, *Volcanology*, 1979.

Volt-ampere

The unit of total apparent power delivered to a load by an alternating current. A circuit branch with E volts across its two terminals, carrying I amperes from one to the other, is said to be receiving EI volt-amperes of apparent power, whatever may be the phase lag θ of current behind voltage. If such a load is driven through a transformer, the volt-amperes into the transformer primary is the same number as the volt-amperes into the load. Other combinations of circuit branches and their volt-amperes (for example, polyphase systems) must be handled carefully, using Eqs. (1) and (2). In Eq. (2), the volt-ampere reactive

$$\begin{aligned}\text{Real power} &= \text{average power delivered to load}\\ &= \text{apparent power} \times \text{power factor}\\ &= EI\cos\theta \quad \text{watts} \qquad (1)\end{aligned}$$

$$\begin{aligned}&\text{Reactive power}\\ &\quad = \text{apparent power} \times \text{reactive power factor}\\ &\quad = EI\sin\theta \quad \text{volt-amperes reactive} \qquad (2)\end{aligned}$$

(var) is 1 volt x 1 ampere when voltage and current are 90° out of phase. The reactive power factor is positive for inductive-reactance loads and negative for capacitive-reactance loads.

Practical computation of real power, reactive power, and apparent power is usually done with complex-number algebra using the geometrical diagram of **Fig. 1**. Clearly Eqs. (3) and (4) hold.

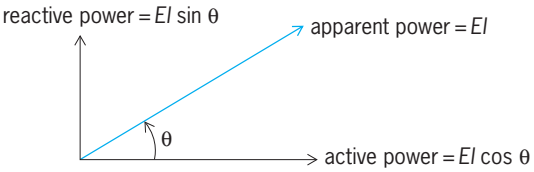


Fig. 1. Apparent, active, and reactive power.

$$\begin{aligned}&(\text{Apparent power})^2\\ &\quad = (\text{real power})^2 + (\text{reactive power})^2 \qquad (3)\\ &(\text{Volt-amperes})^2 = (\text{watts})^2 + (\text{vars})^2 \qquad (4)\end{aligned}$$

See ALTERNATING-CURRENT CIRCUIT THEORY.

Measurement. Alternating-current voltage, current, and real (or active) power can be measured by voltmeters, ammeters, and wattmeters. Reactive power, the average product of instantaneous current and out-of-phase voltage, can be measured with a varmeter (**Fig. 2**), which is in essence a wattmeter movement with provision for adjusting potential-coil current to phase shift of 90° exactly. A special meter to register volt-amperes is seldom used. *See* AMMETER; WATTMETER.

Combinations. Electrical systems of practical interest usually have many branches and possibly several driving generators. The composite voltage-and-current sums given by Eqs. (5) and (6) can be found

$$\sum_{\substack{\text{All}\\ \text{generators}}} (\text{real power in}) = \sum_{\substack{\text{All}\\ \text{branches}}} (\text{real power into branch}) \qquad (5)$$

$$\sum_{\substack{\text{All}\\ \text{generators}}} (\text{reactive power in}) = \sum_{\substack{\text{All}\\ \text{branches}}} (\text{vars into branch}) \qquad (6)$$

for such combinations. These relations are much used by power system engineers, and show that apparent power does not add arithmetically when several circuit branches are combined. The rule of addition for real power, Eq. (5), is a result of the conservation of energy. The rule of addition for reactive power, Eq. (6)—with proper subtraction of

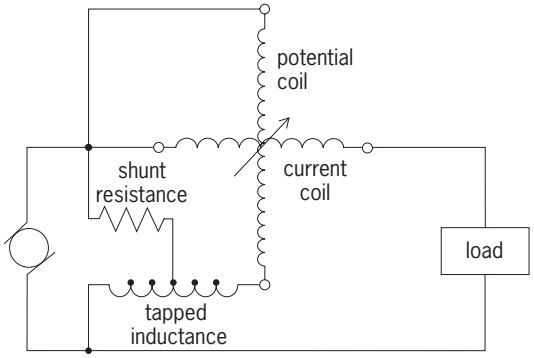


Fig. 2. Single-phase reactive power measurement which uses an inductance.

capacitive vars from inductive vars—is a mathematical statement and must be proved as such.

When the sums of volt-amperes, watts, and volt-amperes reactive are meaningful for multigenerator, multiload systems, the meanings are those obtained from Eqs. (5) and (6). A case of special interest is that of polyphase systems. Real power (watts) always exists, and the composite sums of volt-amperes and vars can be useful if the system is approximately balanced. A badly unbalanced polyphase system is not often described in terms of apparent power. *See* ELECTRIC POWER MEASUREMENT. Mark G. Foster

Bibliography. D. A. Bell, *Electric Circuits*, 6th ed., 1998; L. Chua, C. A. Desoer, and E. Kuh, *Linear and Nonlinear Circuits*, 1987; J. W. Nilsson and S. A. Riedel, *Electric Circuits*, 7th ed., 2004; L. M. Thompson, *Electrical Measurements and Calibration*, 2d ed., 1994.

Voltage amplifier

An electronic circuit whose function is to accept an input voltage and produce a magnified, accurate replica of this voltage as an output voltage. The voltage gain of the amplifier is the amplitude ratio of the output voltage to the input voltage. Often, electronic amplifiers designed to operate in different environments are categorized by criteria other than their voltage gain, even though they are voltage amplifiers in fact. Many specialized circuits are designed to provide voltage amplification. *See* AUDIO AMPLIFIER; CASCODE AMPLIFIER; VIDEO AMPLIFIER.

Voltage amplifiers are distinguished from other categories of amplifiers whose ability to amplify voltages, or lack thereof, is of secondary importance. Amplifiers in other categories usually are designed to deliver power gain (power amplifiers, including push-pull amplifiers) or to isolate one part of a circuit from another (buffers and emitter followers). Power amplifiers may or may not have voltage gain, while buffers and emitter followers generally produce power gain without a corresponding voltage gain. *See* BUFFERS (ELECTRONICS); EMITTER FOLLOWER; POWER AMPLIFIER; PUSH-PULL AMPLIFIER.

Ideal amplifier. A simplified equivalent circuit (**Fig. 1**) can be used to represent a physical voltage source an amplifier, and a resistive load. The voltage source model contains an ideal voltage source and a source resistance. The source resistance, R_S, characterizes the inability of a physical voltage source to maintain its voltage at full value when a load is connected across its terminals, while the load resistance, R_L, represents the physical fact that a load must extract energy from the source.

If the load resistance is connected directly to the source, Ohm's law dictates that the voltage across the load is always less than the ideal source voltage. In fact, the ratio of the voltage to the ideal source voltage equals the ratio of the load resistance to the sum of the load and source resistances. Thus, the load voltage is much less than the ideal source voltage if the source resistance is much greater than the load resistance.

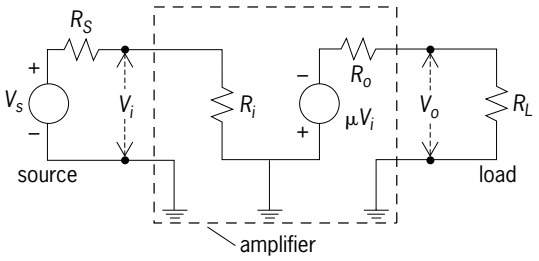


Fig. 1. Voltage amplifier circuit model (with voltage-controlled voltage source). V_s = ideal source voltage, R_S = source resistance, V_i = input voltage, R_i = input resistance, μV_i = voltage-controlled voltage source, R_o = output resistance, V_o = output voltage, R_L = load resistance.

In the circuit model under consideration (Fig. 1), the amplifier is characterized as a black box with an input resistance, R_i; a voltage V_i across the input terminals; a voltage controlled voltage source, μV_i; and an output resistance, R_o. The input resistance of the amplifier acts as a load on the voltage source, and the output of the amplifier creates a new ideal voltage source, μV_i, connected to the load through an output resistance of R_o ohms. To correspond to the convention applied to transistor amplifier modeling, the polarity of this controlled, voltage source is assumed negative with respect to ground. Again, equations can be derived from Ohm's law relating the input voltage to the ideal source voltage, and relating the load voltage to the voltage-controlled voltage source. By combining these equations, the voltage gain of the amplifier, that is, the ratio of the load voltage to the ideal source voltage, is found to be given by Eq. (1).

$$G = -\frac{R_i}{R_i + R_s}\frac{R_L}{R_L + R_o}\mu \qquad (1)$$

If the input resistance is much greater than the source resistance and the output resistance is much less than the load resistance, the magnitude of the voltage gain of the amplifier is approximately μ, where μ typically ranges from one to several million depending on the application and on the internal design of the amplifier. In some cases, the open-circuit voltage of the source is adequately high, and it is necessary only for the amplifier to offset the effect of the load resistance on the source by interposing a unity-gain ($\mu = 1$) voltage amplifier with high input resistance (much greater than the source resistance) and low output resistance (much less than the load resistance) between the source and the load. A voltage amplifier with these characteristics is known as a buffer amplifier.

Alternative amplifier models. Other equivalent circuits are often used to represent voltage amplifiers (**Fig. 2**). For situations where the amplifier input and output resistances are relatively high, the ideal voltage-controlled voltage source of the circuit model discussed above (Fig. 1) is replaced by a voltage-controlled current source, $g_m V_i$, shunted by

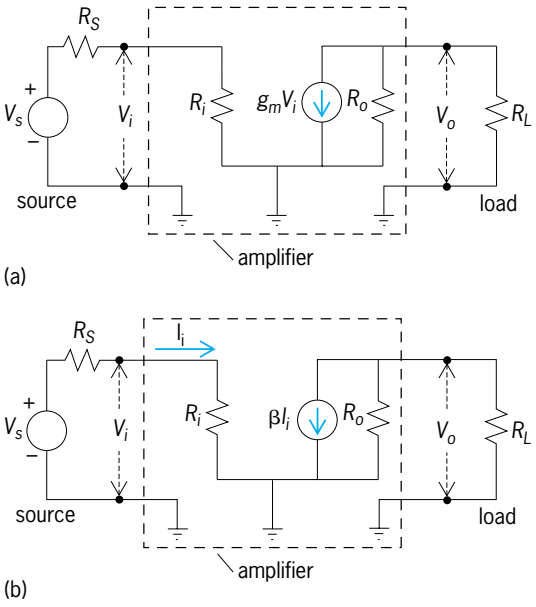


Fig. 2. Alternative voltage amplifier circuit models. (*a*) Model with voltage-controlled current source, $g_m V_i$. (*b*) Model with current-controlled current source, βI_i, where I_i = input current. Other symbols are as in Fig. 1.

the output resistance, R_o (Fig. 2*a*). Exact equivalence between the two circuits is obtained if g_m equals μ divided by the output resistance, and the amplifier voltage gain (for input resistance much greater than source resistance and output resistance much greater than load resistance) can be approximated as the negative of the product of g_m and the load resistance.

A third equivalent circuit model (Fig. 2*b*) for the amplifier is used primarily where the amplifier input resistance is relatively low and its output resistance is high. This circuit uses a current-controlled current source, βI_i, in its output, where I_i is the input current flowing into the amplifier. The model is equivalent to the voltage-controlled current-source model (Fig. 2*a*) if β equals the product of g_m and the input resistance. Here the amplifier gain (for source resistance much greater than input resistance and output resistance much greater than load resistance) can be approximated as the negative of beta times the load resistance divided by the source resistance.

The two current-source models (Fig. 2*a* and *b*) closely resemble the small-signal equivalent circuits used to model single-transistor voltage amplifier circuits: the voltage-controlled current-source model (Fig. 2*a*) usually is used to model a junction field-effect transistor (JFET) amplifier, while the current-controlled current-source model (Fig. 2*b*) is used for the bipolar junction transistor (BJT) amplifier (**Fig. 3**). *See* TRANSISTOR.

Transistor biasing. Neither the field-effect transistor nor the bipolar junction transistor amplifier will operate properly without proper gate (JFET) or base (BJT) bias voltages applied in series with the signal voltage. These bias circuits can be modeled as ideal voltage sources (Fig. 3). The bias and signal voltages are chosen so that the total input voltage—bias plus signal—will not cut off or saturate the amplifier for any value in the range of the input signal voltage. Cutoff will occur, for example, if the base-to-emitter voltage in a bipolar junction transistor drops below about 0.6 V. At cutoff, the transistor output current drops to zero and remains there until the base-to-emitter voltage again increases above the cutoff value. Similarly, if the bipolar junction transistor base-to-emitter voltage is too high, the transistor will saturate and its output voltage will remain steady at a voltage close to zero until the input voltage drops below the saturation level.

In addition to a bias voltage source, well-designed bipolar transistor amplifiers (Fig. 3*b*) require negative feedback at dc to protect the transistor from thermal runaway. The transistor dissipates power in operation, and this power dissipation manifests itself as an increase in the transistor temperature. The transistor collector current is a sensitive temperature-dependent function of its base-to-emitter bias voltage and will rapidly increase with increasing temperature; it takes less bias voltage to obtain the desired collector current as the temperature of the transistor increases. The increasing current further increases the transistor temperature, and this positive feedback process continues until the transistor saturates. Placing a resistor in the emitter circuit causes the base-to-emitter voltage to decrease as the collector current increases, limiting the increase in collector current with increasing temperature. With sufficient negative feedback, this variation of collector current with temperature is controlled easily: a capacitor in the emitter circuit acts as a short circuit at the amplifier operating frequencies, preventing the negative

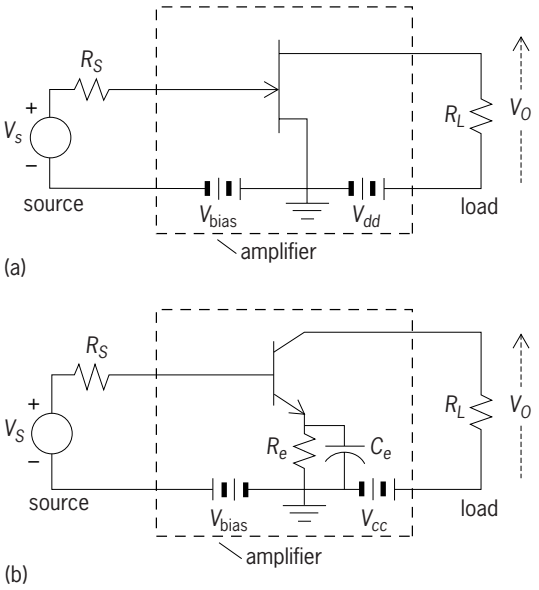


Fig. 3. Basic single-transistor voltage amplifier circuits. (*a*) An *n*-channel junction field-effect transistor common-source amplifier. V_{bias} = bias voltage, V_{dd} = drain supply voltage. (*b*) Bipolar transistor common-emitter amplifier. R_e = emitter resistor, C_e = emitter capacitor, V_{cc} = collector supply voltage. Other symbols are as in Fig. 1.

feedback resistor from reducing the amplifier gain. *See* BIAS (ELECTRONICS).

Interstage coupling. To obtain high gain, cascades of single amplifier circuits are used, usually with a coupling network, actually a simple filter, inserted between the stages of amplification. One such filter is a high-pass network formed by a coupling capacitor, the output resistances of the driving stage, and the input resistance of the driven stage. Since dc voltages are blocked by the capacitor, this ac coupling permits independently setting dc bias voltages for each amplifier stage in the cascade. The coupling network also rejects signals with ac frequency components below a cutoff. The capacitor must be sufficiently large not to attenuate any of the frequencies that are to be amplified. If dc is to be amplified, a direct-coupled amplifier is required, and the design is somewhat more complicated since dc bias voltages on each transistor now cannot be set independently. *See* DIRECT-COUPLED AMPLIFIER; ELECTRIC FILTER.

High-frequency transistor models. The simple models used above for voltage amplifiers and the closely related transistor models are valid only at relatively low operating frequencies. In higher-frequency applications, the parasitic (undesired) capacitance that exists between transistor elements, and sometimes even the small inductance of the transistor leads, must be taken into account. More accurate transistor models incorporate the most important of these parasitic effects.

In most high-frequency amplification applications, gain is desired usually over only a limited range of frequencies, and band-pass filters are used as the interstage coupling networks. These filters are usually realized as inductance-capacitance circuits chosen to resonate, or to impedance-match, the transistor to its load at the center of the frequency band of interest. The transistor parasitic capacitances, supplemented by external capacitors and inductors, are tuned out by the matching network, and parasitic attenuation effects are minimized quite effectively. *See* INTERMEDIATE-FREQUENCY AMPLIFIER; RADIO-FREQUENCY AMPLIFIER.

Balanced and unbalanced amplification. The amplifiers and their models discussed above are called single-ended amplifiers, since their input and output voltages are referred to a common reference point which by convention is called ground. These single-ended circuits, while satisfactory for most noncritical applications, have several weaknesses which degrade their performance in high-gain, weak-signal applications. Their unbalanced construction and their use of a common ground point for return currents makes them susceptible to noise pick-up. Voltage induced in the signal path when extraneous ground currents from other sources flow through the amplifier input signal ground lines is one source of noise. Another is capacitive or inductive coupling between the input signal lines and nearby current-carrying conductors. These coupling mechanisms create noise voltages at the amplifier inputs, which then are amplified along with the desired voltage. To minimize noise on sensitive signal lines, special balanced differential amplifier circuits are often used in critical amplifier applications. Differential amplifiers are designed to have equal impedances to ground for each side of the signal line and to have an output voltage proportional to the difference of the voltages from each signal line to ground. This symmetry cancels common-mode noise voltages, voltages which tend to appear on each of the signal lines as equal voltages to ground. Proper circuit design, with attention to the symmetry of the input circuit construction, can ensure that the majority of undesired noise pick-up will be common-mode noise and, hence, will be attenuated by the differential amplifier. *See* DIFFERENTIAL AMPLIFIER; INSTRUMENTATION AMPLIFIER.

Single-ended amplifiers require multiple resistor networks and interstage coupling capacitors to establish the proper bias conditions for multistage linear amplification. In contrast, differential amplifiers can be biased by using just a few resistors, and their stages also can be direct coupled, obviating the need for coupling capacitors. Since resistors and capacitors are difficult to realize on integrated circuit chips and since they take up much more valuable chip space than do transistors, balanced differential circuits are used almost exclusively in low-frequency and medium-frequency integrated-circuit amplifiers. Discrete transistor single-ended amplifier circuits dominate only at microwave frequencies where it is difficult to control parasitic effects and to achieve high-gain, stable operation with integrated circuits. *See* INTEGRATED CIRCUITS; MICROWAVE.

Differential amplifier. The differential amplifier circuit, as it is usually constructed on a chip (**Fig. 4**), uses transistors configured as current mirrors to set the proper currents through each amplifier transistor and to serve as high-resistance active loads. The pair of transistors that form the first mirror set the total

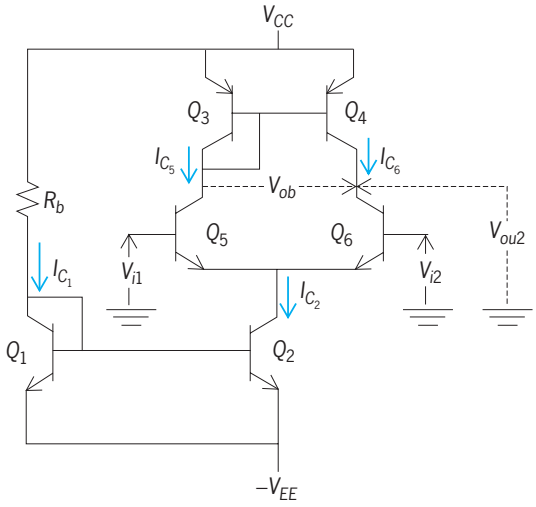


Fig. 4. Bipolar, integrated-circuit differential amplifier. Transistors Q_1, Q_2 form the first current mirror; Q_3, Q_4 form the second current mirror; Q_5, Q_6 are amplifier transistors; V_{i1} and V_{i2} are voltages applied to the input terminals; V_{ou2} is unbalanced output voltage between the collector of Q_6 and ground; V_{ob} is balanced output between collectors Q_5 and Q_6; V_{CC} and V_{EE} are supply voltages; currents $I_{C1} = I_{C2} = I_{C5} = I_{C6} = (V_{CC} + V_{EE} - 0.6)/R_B$.

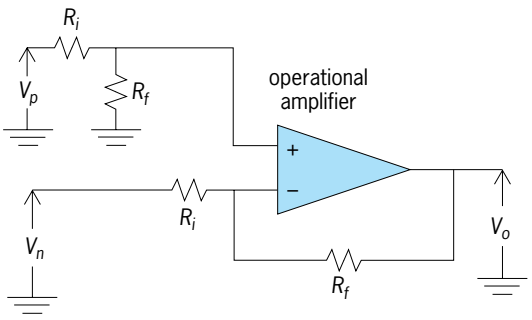


Fig. 5. Operational-amplifier-based difference amplifier, with input voltages V_p and V_n, output voltage V_o, two input resistors R_i, and two other resistors R_f.

collector current through the pair of amplifier transistors. To the signals being amplified, the second current-mirror transistor also appears as a very high resistance in the common emitter lead of the amplifier transistors. *See* CURRENT SOURCES AND MIRRORS.

The second current mirror equalizes the current through the amplifier transistors and also acts as a very high output load resistance on each amplifier transistor. The gain of this single differential stage will be very high if its output voltage is sensed by a circuit with a correspondingly high input resistance.

The differential amplifier can be used either with balanced inputs (with voltages applied to both the input terminals) or with unbalanced ones (where either of these voltages is set to zero by grounding the appropriate terminal). Similarly, the amplifier output voltage will be unbalanced with respect to ground when it is taken between the collector of either of the amplifier transistors and ground, and will be balanced when the output is taken from across the two collectors.

In most applications, needed amplification is incorporated directly into integrated circuits that perform more complex system functions. A single integrated circuit, for example, may contain all of the active (transistor) circuitry needed to build a complete frequency-modulation (FM) radio receiver. Included in the circuit functions contained on the chip are a radio-frequency amplifier, an intermediate-frequency amplifier, and a audio amplifier. The integrated circuit, a handful of discrete components, and a power supply make up the radio. *See* RADIO RECEIVER.

Operational amplifier. In cases where a voltage amplifier is required for some special purpose, operational amplifiers are often used to fill the need. The operational amplifier is an integrated circuit containing a cascade of differential amplifier stages, usually followed by a push-pull amplifier acting as a buffer. The differential voltage gain of the operational amplifier is very high, about 100,000 at low frequencies, while its input impedance is in the megohm range and its output impedance is usually under 100 ohms. The amplifier is designed to be used in a negative-feedback configuration, where the desired gain is controlled by a resistive voltage divider feeding a fraction of the output voltage to the inverting input of the operational amplifier.

A difference amplifier can be created from an operational amplifier and a few resistors (**Fig. 5**). The amplifier both amplifies and subtracts its two input voltages, V_p and V_n. In the case that the two input resistors have the same value, R_i, and the two other resistors in the circuit also have the same value, R_f, Eq. (2) holds for the output voltage, V_o.

$$V_o = \frac{R_f}{R_i}(V_p - V_n) \tag{2}$$

With needed amplification built into many integrated circuits and with the availability of operational amplifiers for special-purpose amplification needs, there is seldom a need to design and build a voltage amplifier from discrete components. *See* AMPLIFIER; OPERATIONAL AMPLIFIER.

Philip V. Lopresti

Bibliography. R. L. Geiger, P. E. Allen, and N. R. Strader, *VLSI Design Techniques for Analog and Digital Circuits*, 1990; P. Gray and R. Meyer, *Analysis and Design of Analog Integrated Circuits*, 4th ed. 2001; P. Horowitz and W. Hill, *The Art of Electronics*, 2d ed., 1989; A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, 4th ed., 1997.

Voltage measurement

Determination of the difference in electrostatic potential between two points. The unit of voltage in the International System of Units (SI) is the volt, defined as the potential difference between two points of a conducting wire carrying a constant current of 1 ampere when the power dissipated between these two points is equal to 1 watt.

Voltage standards. These comprise Josephson junction standards, Weston standard cells, and electronic standards.

Josephson junction standards. The discovery of the Josephson effect in superconducting tunnel junctions in 1962 made possible the operation of a quantum standard of voltage. By 1972 this standard was in widespread use in national standards laboratories to supervise their Weston-cell working standards. The resolution and stability achieved by the quantum standard (1 in 10^9) were better by far than the uncertainty to which its relation to the SI unit of voltage was known (1 in 10^5). The need for some international agreement at a higher level of accuracy was resolved by defining the properties of the Josephson junction as a frequency-to-voltage converter. Although it was recognized at the time that the value chosen was somewhat arbitrary, it became possible to have international standards of adequate precision to support trade in precision electrical measuring instruments.

Subsequent work on the realization of the electrical units has enabled the relationship between the SI unit of voltage and the standard based on the Josephson effect to be established with an uncertainty believed to be slightly better than 1 in 10^7. As a result, the International Committee for Weights and Measures (CIPM; the international body concerned with measurement standards) recommended

a new value for the Josephson constant k_{J-90} given by Eq. (1),which would be used throughout the world

$$k_{J-90} = 483{,}597.9 \text{ GHz/V} \quad (1)$$

from January 1, 1990, onward. As a result the values assigned to the working standards of various laboratories were changed on that date, and all subsequent calibrations carried out by them take account of this. The new value was chosen to be consistent with the SI unit within the limits of current knowledge.

A Josephson junction consists of two superconductors, such as lead or niobium, separated by a thin insulating or poorly conducting barrier across which conduction can occur by the tunneling process. The tunnel barrier may be formed by controlled oxidation or evaporation of a material such as silicon oxide. The dimensions of the junction have to be considerably less than 0.04 in. (1 mm). When such a junction is irradiated by radio-frequency energy, usually in the frequency range 10–100 GHz, its voltage–current characteristic becomes broken into a stepped form. There is considerable hysteresis and overlapping of the steps (**Fig. 1**). The height V of each voltage step is given by Eq. (2), where f is the

$$V = \frac{h}{2e} f \quad (2)$$

irradiating frequency, h is Planck's constant, and e is the electron charge. This voltage can be measured across a single junction. Since the junction must be immersed in a helium cryostat at a temperature of a few kelvins in order to produce the superconducting effect, it is convenient to include a large part of the measuring system (**Fig. 2**) in the cryostat. The ratio between the standard cell and the Josephson step voltages can be measured in terms of the ratio of two resistors in the system (Fig. 2), a ratio which has to be established by a separate experiment. Another Josephson junction can be used as a null detector, and hence the step voltage of a few millivolts can be compared with that of a room-temperature standard cell to within less than 1 part in 10^8. *See* SQUID.

By using photolithographic techniques, it has proved possible to construct arrays of Josephson

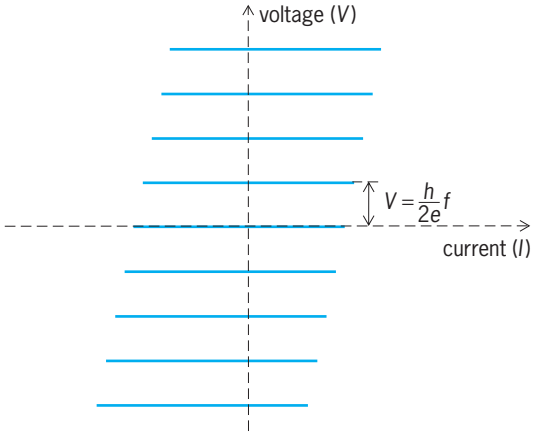


Fig. 1. Voltage–current characteristic of ideal Josephson junction.

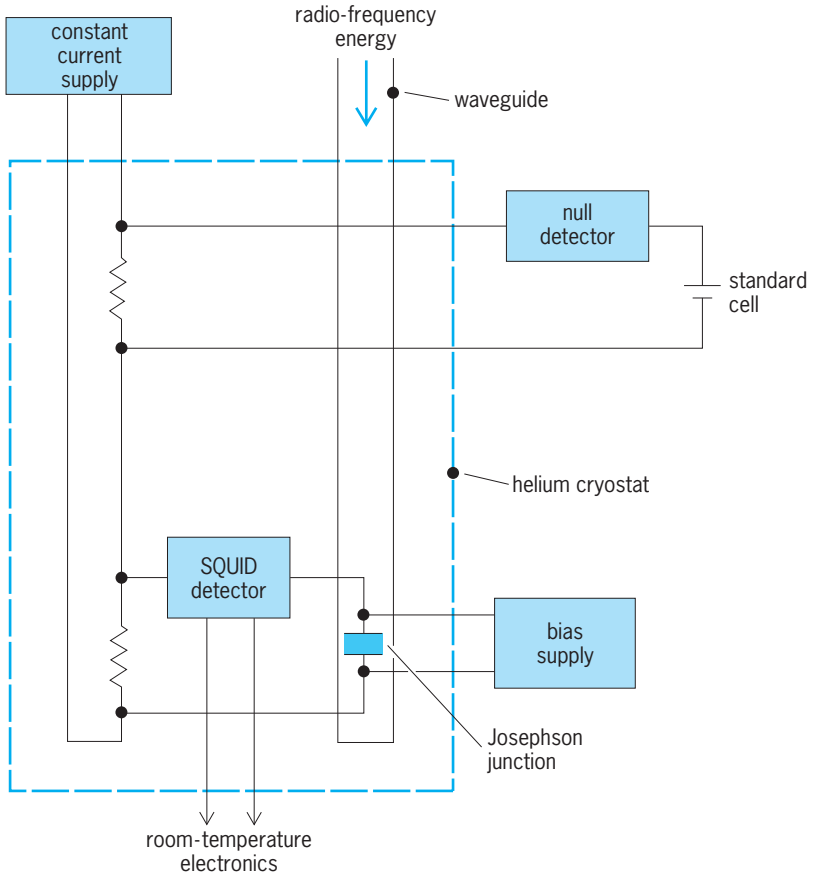


Fig. 2. Simpified Josephson junction measuring system.

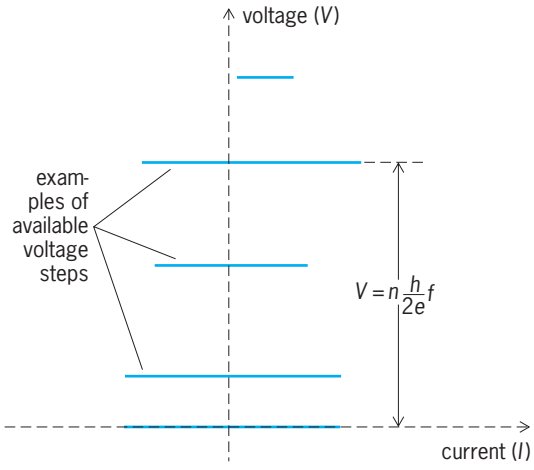


Fig. 3. Voltage–current characteristic of Josephson junction array.

junctions that consist of thousands of separate devices that can function simultaneously. In such an array each device can operate at a relatively high voltage step that extends across the zero-current axis (**Fig. 3**), and so the array may operate with a voltage of over 10 V across it. By using such a system, measurements can be made directly against other reference devices with a precision limited only by their stability and noise properties. *See* JOSEPHSON EFFECT; SUPERCONDUCTING DEVICES.

Standard cells. The saturated Weston cell, which consists of electrodes of mercury and mercury-cadmium amalgam in an electrolyte of saturated cadmium sulfate solution, has been in use since the latter part of the nineteenth century as a portable room-temperature voltage standard. *See* ELECTROMOTIVE FORCE (CELLS).

Although many designs of digital voltmeter are unsuitable for use with Weston cells as a result of excessive input currents, it is possible to choose a suitable model. By using one of these or a relatively simple potentiometer, standard-cell voltages can be compared to better than 1 part in 10^7. By comparing the voltage of each cell against the mean of a group it is possible to detect and eliminate less stable cells, and thus to maintain a voltage standard of high stability. Saturated standard cells are very sensitive to environmental conditions and to the passage of any current through them, and are therefore rather unsatisfactory as reference standards and even more so as transfer standards. Nevertheless, the stability of good cells may be better than 100 nanovolts per year. *See* POTENTIOMETER.

Electronic voltage standards. Diodes used for voltage regulation (Zener diodes) pass a current that increases rapidly when the voltage rises above some critical value, and hence tend to act as constant voltage devices. If such diodes are fed from a well-stabilized source, for example, if they are used in an operational amplifier configuration, they can provide voltage references that may have a stability as good as 1 in 10^6. The temperature coefficient of simple Zener diodes is a function of voltage and operating current. This coefficient can be greatly reduced by the use of compensated diodes, in which one or more forward-biased junctions (having a negative temperature coefficient of voltage) are connected in series with a Zener breakdown junction chosen to have a positive coefficient. A further possibility is the combination of transistor and Zener junctions to provide a compensated reference element. Reference-quality diodes can also be constructed as part of an integrated circuit that includes built-in temperature stabilization (**Fig. 4**). External operational amplifiers and other components are used to complete the voltage and temperature regulation circuits. Like Weston cells, the best diodes are obtained by selection. Since they are less sensitive to environmental conditions and overloads than standard cells, they are much more suitable for use as transfer standards and are also in almost universal use as voltage references in electronic measuring instruments. *See* INTEGRATED CIRCUITS; OPERATIONAL AMPLIFIER; ZENER DIODE.

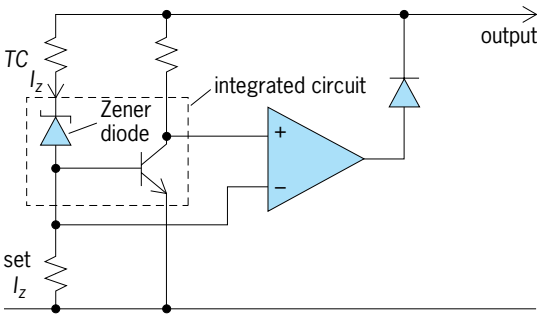


Fig. 4. Reference circuit using an integrated reference amplifier. The resistor "set I_z" allows current I_z flowing through the Zener diode to be adjusted independently of the transistor current. The resistor *TC* allows the temperature coefficient of the output voltage to be adjusted.

Realization of the unit of voltage. Reference standards of voltage are obtained through the conventional value of k_{J-90} recommended by the CIPM, the Josephson effect, and a measurement of frequency. Accurate realization of the SI unit involves exceedingly difficult measurements, for example, combining a realization of the ampere by using a current balance with the unit of resistance obtained through a calculable capacitor. For a full discussion of the unit and its relationship to other electrical units. *See* ELECTRICAL UNITS AND STANDARDS.

Direct-current voltage measurement. The chief types of instruments for measuring direct-current (constant) voltage are potentiometers, resistive voltage dividers, pointer instruments, and electronic voltmeters.

Potentiometer measurements. The most fundamental dc voltage measurements from 0 to a little over 10 V can now be made by direct comparison against Josephson systems. At a slightly lower accuracy level and in the range 0 to 2 V, precision potentiometers are used, in conjunction with very low-noise electronic amplifiers or photocoupled galvanometer detectors. Potentiometers are capable of self-calibration, since only linearity is important, and can give accurate measurements down to a few nanovolts. When electronic amplifiers are used, it may often be more convenient to measure small residual unbalance voltages, rather than to seek an exact balance. *See* AMPLIFIER; GALVANOMETER.

In all dc low-voltage measurements, care is necessary to avoid thermal emf's. All wiring must be of pure copper, which may be plated with gold but not tin or nickel, and any movement of conductors must be avoided. It is necessary to wait several seconds after any switch operation to allow thermal emf's that are generated to die away.

Resistive voltage dividers. Voltage measurements of voltages above 2 V are made by using resistive dividers. These are tapped chains of wire-wound resistors, often immersed in oil, which can be self-calibrated for linearity by using a buildup method. Instruments for use up to 1 kV, with tappings typically in a binary or binary-coded decimal series from 1 V, are known as volt ratio boxes, and normally provide uncertainties down to a few parts per million. Another configuration allows the equalization of a string of resistors, all operating at their appropriate power level, by means of an internal bridge. The use of series-parallel arrangements can provide certain easily adjusted ratios (**Fig. 5**).

Higher voltages can be measured by extending such chains, but as the voltage increases above about 15 kV, increasing attention must be paid to avoid any sharp edges or corners, which could give rise

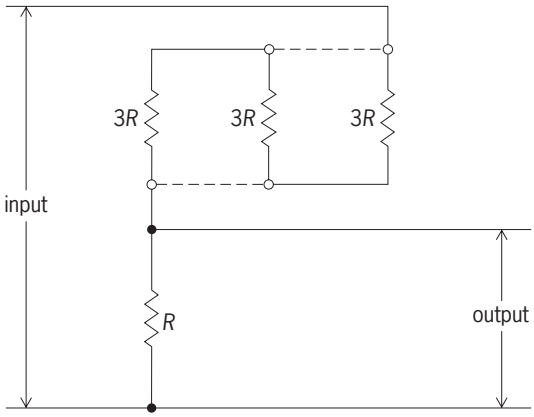


Fig. 5. Achievement of a 10:1 resistance ratio by a series-parallel connection. The single resistor *R* is adjusted to be equal in value to the three resistors of value 3*R* connected in parallel. When all the resistors are reconnected in series, a ratio of exactly 10 is produced.

to corona discharges or breakdown. High-voltage dividers for use up to 100 kV with an uncertainty of about 1 in 10^5, and to 1 MV with an uncertainty of about 1 in 10^4, have been made. *See* CORONA DISCHARGE; ELECTRICAL BREAKDOWN.

DC pointer instruments. For most of the twentieth century the principal dc indicating voltmeters have been moving-coil milliammeters, usually giving full-scale deflection with a current between 20 microamperes and 1 milliampere and provided with a suitable series resistor. Many of these will certainly continue to be used for many years, giving an uncertainty of about 1% of full-scale deflection.

Electronic voltmeters. Since about 1950 it has become increasingly common to use electronic amplifiers to drive pointer instruments in order to reduce the loading of the measured signal, reduce the possibility of mechanical damage to the meter through overload, and provide greatly increased sensitivity. It is possible to obtain instruments having a full-scale deflection of less than 1 nanovolt, and such instruments are now widely used in laboratory work as null detectors.

A further development, the replacement of the analog indicator by a digital display, produced the digital voltmeter (DVM). This type has now become the principal means used for voltage measurement at all levels of accuracy, even beyond one part in 10^7, and at all voltages up to 1 kV. Essentially, digital voltmeters consist of a power supply, which may be fed by either mains or batteries; a voltage reference, usually provided by a Zener diode; an analog-to-digital converter; and a digital display system. This design provides measurement over a basic range from zero to a few volts, or up to 20 V. Additional lower ranges may be provided by amplifiers, and higher ranges by resistive attenuators. The accuracy on the basic range is limited to that of the analog-to-digital converter. *See* ANALOG-TO-DIGITAL CONVERTER; ELECTRONIC POWER SUPPLY.

Most modern digital voltmeters use an analog-to-digital converter based on a version of the charge balance principle. In such converters the charge accumulated from the input signal during a fixed time by an integrator is balanced by a reference current of opposite polarity. This current is applied for the time necessary to reach charge balance, which is proportional to the input signal. The time is measured by counting clock pulses, suitably scaled and displayed. Microprocessors are used extensively in these instruments. *See* MICROPROCESSOR.

Alternating-current voltage measurements. The mean voltage of any number of complete cycles of an ac waveform is always zero. Any departure from zero indicates some dc content. By its nature, an ac signal is always changing. It is therefore necessary to define what is being measured in alternating voltage measurements. The value of greatest interest is the root-mean-square (rms) value, since this is a direct measure of the energy content. The root-mean-square value V_{rms} is defined in terms of the instantaneous value V by Eq. (3). The rectified mean

$$V_{\text{rms}} = \sqrt{\frac{1}{T}\int_0^T V^2 \underset{\lim T\to\infty}{dt}} \tag{3}$$

value is also important since dc instruments fitted with rectifiers usually respond to the mean of the modulus of the signal, often (incorrectly) referred to as the mean. One definition of V_{mean} is given by Eq. (4).

$$V_{\text{mean}} = \frac{1}{T}\int_0^T |V| \underset{\lim T\to\infty}{dt} \tag{4}$$

In any waveform there is a fixed relationship between the peak, root-mean-square, and rectified mean values. For a sinusoidal waveform, the root-mean-square and peak values are related by Eq. (5), and the root-mean-square and rectified mean by Eq. (6).

$$V_{\text{rms}} = \frac{1}{\sqrt{2}} V_{\text{peak}} \tag{5}$$

$$V_{\text{rms}} = \frac{\pi}{2\sqrt{2}} V_{\text{mean}} \tag{6}$$

It has become conventional for instrument scales to be marked with the root-mean-square value that a sine wave would have, if it gave the same deflection as the actual voltage applied. Mean sensing instruments thus have a built-in scaling factor of $1.1107198 = \pi/(2\sqrt{2})$. An alternating voltage is rarely purely sinusoidal, and departure from a sine form is usually specified in terms of harmonic content. *See* NONSINUSOIDAL WAVEFORM; WAVEFORM; WAVEFORM DETERMINATION.

Thermal transfer. Since the working standards of voltage are of the direct-current type, all ac measurements have to be referred to dc through transfer devices or conversion systems. Any instrument that responds to both ac and dc voltage can be used as a transfer device, but the transfer errors may need to be determined. The most accurate transfer device is the multijunction thermal converter (MJTC). This has a calculated ac-dc difference well below 1 part per million over its useful frequency range of about

30 Hz to 10 kHz, and the good agreement between different devices of this type supports this value. These devices are used to select and calibrate single-junction converters, which are more readily available, have only slightly inferior performance, and are useful from 10 Hz to hundreds of megahertz.

Single-junction thermal converters (SJTCs) consist of a thin wire, usually of nickel-chrome alloy, which is heated by the incoming signal to produce a rise in temperature of about 250°F (140°C). The temperature rise is detected by a single thermo- couple junction fixed in an insulating bead near the center of the wire. The whole device is enclosed in an evacuated glass envelope. The heating current is usually 5 or 10 mA. The device is calibrated by comparison with a similar device whose performance has already been established or a multijunction converter, by the sequential application of ac, dc, reversed dc, and ac signals. The pairs of results for ac and dc are averaged to eliminate the effect of drift and current reversal. More complex sequences can be used to obtain further improvements. By this means the ac-dc difference of the best devices can be certified to a level of about 5 parts per million (ppm) at low frequencies, degrading with frequency to about 15 ppm at 1 MHz.

In commercial multirange instruments, similar single-junction thermal converters are combined with switched series resistors of special design to give voltage ranges from 1 V to about 1 kV. The measurement at the highest level of accuracy of an alternating voltage is thus made by establishing a dc voltage equivalent to the root-mean-square ac voltage through the use of such an instrument. Thermal converters are also made in which other methods of temperature sensing are used, for example, the effect on the base-to-emitter voltage of a planar transistor. *See* THERMAL CONVERTERS; THERMOCOUPLE; TRANSISTOR.

AC conversion. A variety of techniques can be used to convert an ac signal into a dc equivalent automatically. All multimeters and most ac meters make use of ac-dc conversion to provide ac ranges. These are usually based on electronic circuits. Rectifiers provide the most simple example. *See* MULTIMETER.

In a commonly used system (**Fig. 6**), the signal to be measured is applied, through a relay contact, to a thermal converter. In order to improve sensitivity, a modified single-junction thermal converter may be used in which there are two or three elements in a single package, each with its own thermocouple. The output of the thermal converter is measured by a very sensitive, high-resolution analog-to-digital converter, and the digital value memorized. When a measurement is required, the relay is operated, and the thermal converter receives its input, through a different relay contact, from a dc power supply, the amplitude of which is controlled by a digital and analog feedback loop in order to bring the analog-to-digital converter output back to the memorized level. The dc signal is a converted value of the ac input and can be measured. Modern versions of this type of instrument make use of microprocessors to control the conversion process, enhance the speed of operation, and include corrections for some of the errors in the device and range-setting components.

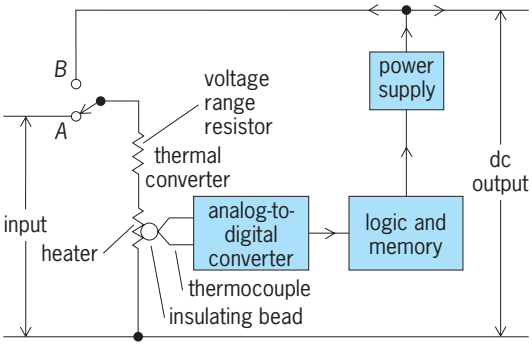


Fig. 6. Thermal ac-dc converter system. The thermal converter receives the signal to be measured through relay contact *A*, and the input from the dc power supply through relay contact *B*.

Voltage ratio devices. Two distinct forms of transformer are used as ratio standards at power and audio frequencies, namely, voltage transformers and inductive voltage dividers.

Voltage transformers have separate primary and secondary windings and operate with the secondary circuit very lightly loaded. High-permeability cores are operated at a low flux density to give good linearity. The primary/secondary voltage ratio then closely approaches the turns ratio, and phase errors are very small. Ratio and phase errors can be measured by bridge techniques using inductive or capactive dividers, and once measured are normally stable through the life of the transformer. A good reference voltage transformer may have errors less than 50 ppm. Voltage transformers are used chiefly at power frequencies and voltages up to 500 kV for calibration purposes. *See* INSTRUMENT TRANSFORMER; TRANSFORMER.

Inductive voltage dividers are basically autotransformers, usually toroidal in form, in which a tapped primary winding is wound on a core of very high permeability, usually of supermumetal. To achieve exact equality of division between windings, the winding consists of a rope of 10 (for a decade divider) similar strands of insulated wire, which are then connected end to end to give the successive sections. This construction provides an equality of division better than 1 ppm. In a two-stage divider the toroidal magnetic core is subdivided, and a second winding arranged to thread only part of it. This second winding is arranged to carry a current that generates sufficient flux in the one core to provide almost all the magnetization for the whole transformer, usually by having the same number of turns as the main winding and being connected to the same source (**Fig. 7**). The equality of division is improved to better than 1 in 10^8 for a single decade of a two-stage divider. Such dividers can be connected in a decade arrangement to give variable ratios with the same resolution, and are used as a building block in precision ac measurements. *See* AUTOTRANSFORMER; INDUCTIVE VOLTAGE DIVIDER.

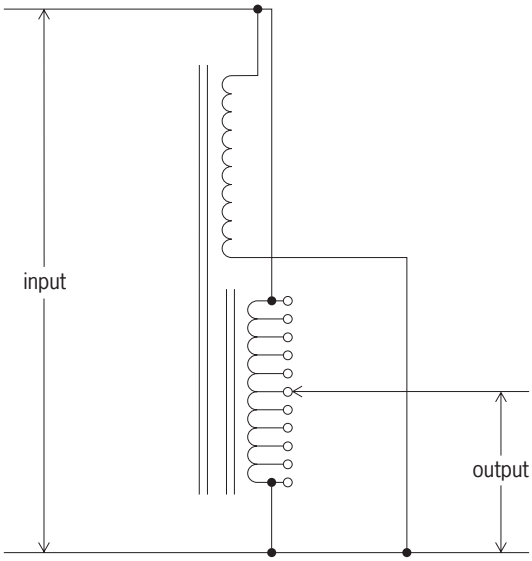


Fig. 7. Two-stage inductive voltage divider transformer.

Digital voltmeters. As in the dc case, digital voltmeters are now probably the instruments in widest use for ac voltage measurement. The simplest use diode rectification of the ac to provide a dc signal, which is then amplified and displayed as in dc instruments. This provides a signal proportional to the rectified mean. For most purposes an arithmetic adjustment is made, and the root-mean-square value of a sinusoidal voltage that would give the same signal is displayed. Several application-specific analog integrated circuits have been developed for use in instruments that are required to respond to the root-mean-square value of the ac input. More refined circuits, based on the logarithmic properties of transistors or the Gilbert analog multiplier circuit, have been developed for use in precision instruments. The best design, in which changes in the gain of the conversion circuit are automatically compensated, achieves errors less than 10 ppm at low and audio frequencies.

Sampling digital voltmeters are also used, in which the applied voltage is switched for a time very short compared with the period of the signal into a sample-and-hold circuit, of which the essential element is a small capacitor. The voltage retained can then be digitized without any need for haste. At low frequencies this approach offers high accuracy and great versatility, since the voltages can be processed or analyzed as desired. At higher frequencies, for example, in the microwave region, it also makes possible the presentation and processing of fast voltage waveforms using conventional circuits. *See* OSCILLOSCOPE.

AC pointer instruments. The analog methods mentioned above produce dc output voltages that are proportional to the ac values, and hence at lower levels of accuracy can be used to drive dc pointer instruments. However, most pointer ac instruments are either of the rectifier type, sensitive to the rectified mean value and calibrated to indicate the root-mean-square value of an equivalent sinusoidal voltage, or of the moving-iron type, which provides a square-law scale. Electrostatic voltmeters, which also produce a square-law scale, are sometimes used for voltages in the range 1–10 kV.

Frequencies above 1 MHz. Voltage measurements at radio frequencies are made by the use of rectifier instruments at frequencies up to a few hundred megahertz, single-junction converters at frequencies up to 500 MHz, or matched bolometers or calorimeters. At these higher frequencies the use of a voltage at a point must be linked to information regarding the transmission system in which it is measured, and most instruments effectively measure the power in a matched transmission line, usually of 50 ohms characteristic impedance, and deduce the voltage from it. *See* BOLOMETER; MICROWAVE MEASUREMENTS; MICROWAVE POWER MEASUREMENT; TRANSMISSION LINES.

Pulse measurements. Pulse voltage measurements are made most simply by transferring the pulse waveform to an oscilloscope, the deflection sensitivity of which can be calibrated by using low-frequency sine waves or dc. Digital sampling techniques may also be used. *See* ELECTRICAL MEASUREMENTS. R. B. D. Knight

Bibliography. C. H. Dix and A. E. Bailey, Electrical standards of measurement, *Proc. IEE*, vol. 122, no. 10R, October 1975; S. Geczy, *Basic Electrical Measurements*, 1984; F. K. Harris, *Electrical Measurements*, 1952, reprint 1975; H. W. Hellwig (ed.), Conference of Precision Electromagnetic Measurements, *IEEE Trans. Instrum. Meas.*, vol. 32, no. 1, March 1983; *International System of Units (SI)*, NBS Spec. Publ. 330, 1991; B. Noltingk (ed.), *Instrumentation Reference Book*, 2d ed., 1995; M. V. Reissland, *Electrical Measurements: Fundamentals, Concepts, Applications*, 1989; M. G. Say (ed.), *Electrical Engineer's Reference Book*, 15th ed., 1993.

Voltage-multiplier circuit

A circuit which produces a dc output voltage that is a fixed multiple of the input dc voltage or the peak voltage of an ac input waveform. This fixed multiple is approximately an integer ($\pm 2, \pm 3, \pm 4, \ldots$). Voltage multipliers are used to produce a high dc voltage where modest load current is required. The circuit can be implemented with diodes and capacitors or with switched capacitors.

Diode-capacitor voltage multipliers. A diode passes conventional current in only one direction. In the diode-capacitor voltage doubler, this current through one of the diodes charges one of the capacitors to approximately the peak value of the input voltage. The series combination of the input source and the capacitor generates a voltage across the diode that varies from approximately 0 to -2 times the peak input voltage. The diode is reverse biased except for short intervals to replace charge drained from the capacitor to supply the load current or leakage. A second diode and capacitor form another peak detecting circuit with the voltage across the first diode as the input. This results in a negative dc output voltage equal to twice the peak of the input waveform. The polarity of the output can be reversed by simply

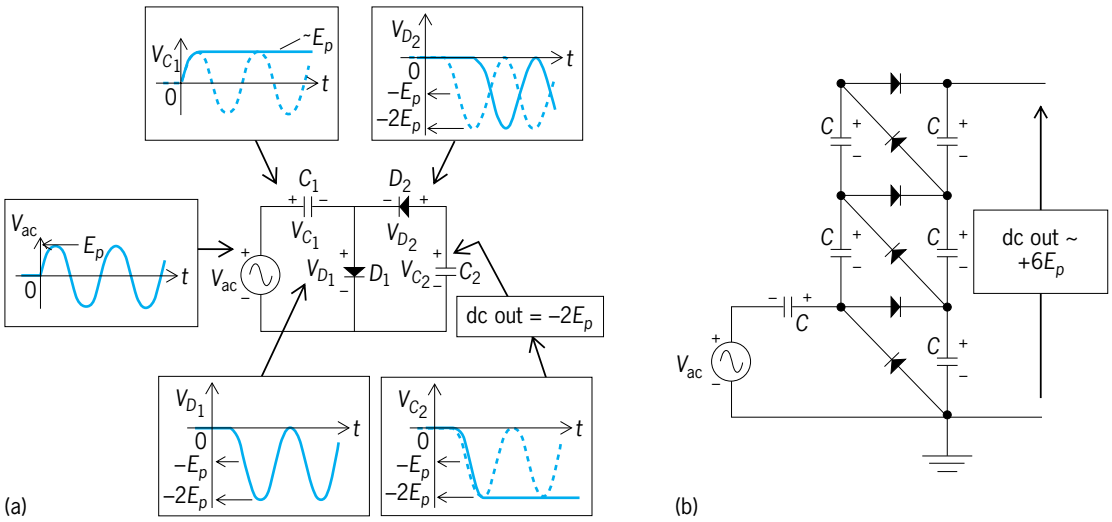


Fig. 1. Diode-capacitor voltage-multiplier circuits. (*a*) Negative voltage doubler. Diodes pass conventional current only in the directions of the arrows. Waveforms of input ac voltage, V_{ac} with peak value E_p, and of voltages at diodes D_1 and D_2 and capacitors C_1 and C_2 are also shown. (*b*) Positive, times-six multiplier with three stages, each with two diodes and two capacitors, C.

reversing both diodes. Another version of the voltage doubler produces two outputs, one equal to the peak input voltage and the other to its negative, using the same components. This is realized by connecting a positive and a negative peak detector across the input. *See* CAPACITOR; DIODE.

A positive times-six voltage multiplier circuit (**Fig. 1***b*) operates on the same principle as the voltage doubler circuit. Each stage, consisting of two diodes and two capacitors, adds two to the multiplication factor. The forward-bias voltage drop across each diode that results when charge is forced on the capacitors will cause the output to be less than that for an ideal. Also, any current required by the load will produce an ac ripple in the dc output. If the first capacitance is much greater than the second in the voltage-doubler circuit discussed above (Fig. 1*a*), the peak-to-peak ripple voltage can be approximated by the equation below, where I_{dc} and E_{dc} are the output

$$\text{Peak-to-peak ripple} = \frac{I_{dc}}{fC_2} = \frac{E_{dc}}{fC_2R_{load}}$$

current and voltage respectively, f is the fundamental frequency of the input ac waveform, C_2 is the second capacitance, and R_{load} is the load voltage. *See* RIPPLE VOLTAGE.

Switched-capacitor applications. A typical voltage-multiplier circuit that uses switched capacitors (**Fig. 2***a*) is a dc-to-dc converter that generates +10-V and −10-V dc sources from a single +5-V dc source. A common application for this circuit is to use a +5-V dc power supply to power integrated circuits which require positive and negative voltages. Such switched-capacitor applications are well suited to integrated circuits because the transistor switches and the associated control circuitry take up very little space and the added cost is negligible. The switches are frequently shown as mechanical switches on circuit diagrams (Fig. 2*b*) for simplicity, but they are actually transistors—typically, parallel positive and negative metal-oxide-semiconductor (PMOS and NMOS) field-effect transistors forming a transmission gate switch. The switches are alternately connected to one set of terminals and then the other. Special nonoverlapping clock signals (Fig. 2*b*) must be provided to ensure that both sets of terminals are never on or connected at the same time. In one switch

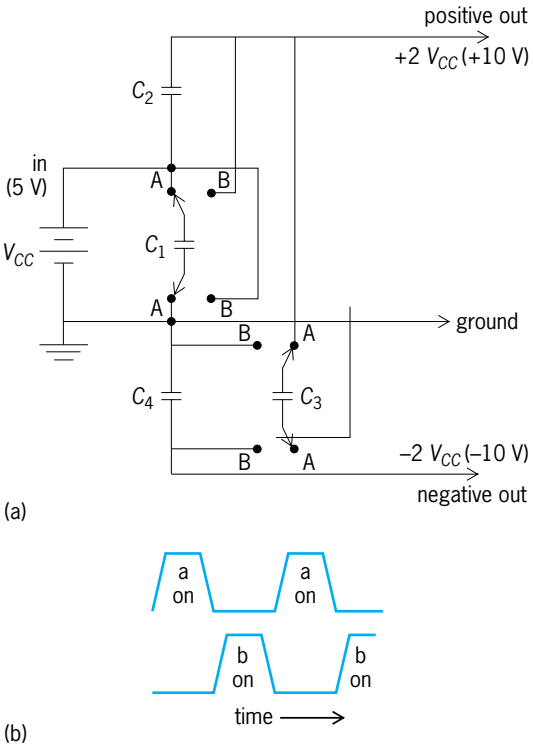


Fig. 2. Dual switched-capacitor dc-to-dc converter that generates +10-V and −10-V dc sources from a single +5-V dc source. (*a*) Simplified circuit diagram. Switches are alternately connected to terminals A and B. C_1, C_2, C_3, and C_4 are capacitors. V_{CC} is the supply voltage. (*b*) Waveforms of nonoverlapping clock signals.

position, one of the capacitors is connected to the input, and it rapidly charges to the supply voltage. In the other switch position, this capacitor is connected to a second capacitor. This is the switched-capacitor equivalent to a low-pass filter, and the second capacitor will charge up to the supply voltage with an equivalent time constant equal to C_2/fC_1, where C_1 and C_2 are the first and second capacitances, and f is the clock frequency. A voltage doubler is realized at the positive output of the circuit. To generate the negative output, a third capacitor is first connected across the positive output and then across a fourth capacitor. *See* ELECTRONIC POWER SUPPLY; INTEGRATED CIRCUITS; SWITCHED CAPACITOR; TRANSISTOR.

Applications using diode multipliers. Voltage-multiplier circuits were historically used to generate the high voltage required for cathode-ray tubes and photomultipliers, whose current requirements are very small. The voltage-doubler circuit discussed above (Fig. 1*a*) can be used as a dc-to-dc converter to generate a negative-bias voltage for an integrated circuit that has only a positive supply voltage. The ac voltage source is replaced with an internal ring or relaxation oscillator that generates a square wave, for example, 0 to 5 V dc. The voltage across the first capacitor is less than the dc supply voltage because of the forward-bias voltage drop across the diode. The ac voltage across the first diode will be a square wave ranging from approximately the negative of the supply voltage to 0. The second diode and capacitor detect the peak negative voltage, which is approximately −3.2 V dc for a +5-V dc supply. This voltage is adequate to provide the necessary negative bias required by some integrated circuits. This application of a voltage doubler circuit is called a substrate pump or a substrate bias generator. *See* CATHODE-RAY TUBE; PHOTOMULTIPLIER.

Norman G. Dillman

Bibliography. S. G. Burns and P. R. Bond, *Principles of Electronic Circuits*, 2d ed., 1996; R. C. Dorf (ed.), *The Electrical Engineering Handbook*, 2d ed., 1997; E. D. Fabricius, *Introduction to VLSI Design*, 1990; C. A. Schuler, *Electronics: Principles and Applications*, 5th ed., 1999.

Voltage regulation

The change in voltage magnitude that occurs when the load (at a specified power factor) is reduced from the rated or nominal value to zero, with no intentional manual readjustment of any voltage control, expressed in percent of nominal full-load voltage. Voltage regulation is a convenient measure of the sensitivity of a device to changes in loading. This concept is commonly used for transmission or distribution circuits as well as power supply equipment, including constant-voltage transformers, rectifiers, outdoor coupling capacitors, direct-current generators, induction frequency converters, synchronous generators, line regulation circuits, and thyristor converters. A common form of expressing voltage regulation in percent is shown in the equation below.

$$\text{Voltage regulation} = \frac{\text{no-load voltage} - \text{full-load voltage}}{\text{full-load voltage}} \times 100$$

As an example of the use of this equation, one can compute the regulation of the supply voltage to an appliance, home, office, or an entire city. This gives a measure of the need for special voltage-regulating equipment, particularly if there are voltage-sensitive components in the load. *See* GENERATOR; TRANSFORMER; VOLTAGE REGULATOR.

Paul M. Anderson

Bibliography. D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 2000; B. R. Gungor, *Power Systems*, 1988; F. Jay (ed.), *IEEE Standard Dictionary of Electrical and Electronic Terms*, IEEE Std100-1988, 4th ed., 1988.

Voltage regulator

A device or circuit that maintains a load voltage nearly constant over a range of variations of input voltage and load current. Voltage regulators are used wherever the unregulated voltage would vary more than can be tolerated by the electrical equipment using that voltage. Alternating-current distribution feeders use regulators to keep the voltage supplied to the user within a prescribed range. Electronic equipment often has voltage regulators in dc power supplies.

Electronic Voltage Regulators

A dc power supply is an essential component in any electronic system. **Figure 1** shows a Thévenin equivalent circuit of a power supply voltage source having an open-circuit (no-load) voltage V_0 and output resistance R_0. When a load with resistance R_L is connected to the output of this supply, a load current $I_L = V_L/R_L$ flows through R_0, resulting in a drop in load voltage V_L as given by Eq. (1).

$$V_L = V_0 - I_L R_0 \tag{1}$$

See THÉVENIN'S THEOREM (ELECTRIC NETWORKS).

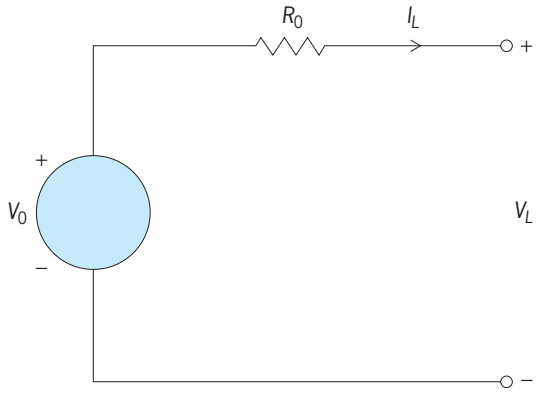


Fig. 1. Equivalent circuit of a basic dc power supply.

A basic power supply as modeled above exhibits two undersirable characteristics. The first is that the load voltage V_L decreases with increasing load current I_L. This effect can be severe in supplies with large effective output resistances. Such supplies are said to have poor load regulation. The second problem is that V_L depends directly on the source V_0. In practice, V_0 might be derived from a relatively inaccurate source such as the ac line voltage (after suitable rectification and filtering) or a battery. Fluctuations in V_0 reflect directly onto the voltage experienced by the load. In this case, the supply is said to have poor line regulation. An ideal power supply would exhibit perfect load regulation (V_L independent of I_L) and perfect line regulation (V_L independent of V_0) and would provide an output voltage of the form of Eq. (2), where V_{ref} is a well-defined reference voltage and k is a constant scaling factor. The task of an electronic voltage regulator is to provide an output voltage characteristic that closely approximates the ideal of Eq. (2), given an unregulated supply voltage as an input.

$$V_L = kV_{\text{ref}} \tag{2}$$

Feedback network. Typically, electronic voltage regulators employ a feedback network of the form shown in **Fig. 2**. Here, a high-gain amplifier compares a fraction of the load voltage V_L/k with a constant reference V_{ref}. Any difference between these two voltages is amplified and used to control a series pass device in a manner whereby this difference is minimized. For an ideal amplifier with zero offset and infinite voltage gain, the difference is reduced to zero and the ideal relationship of Eq. (2) is realized. *See* FEEDBACK CIRCUIT.

Monolithic voltage regulator chips. The wide range of applications for electronic voltage regulators has led to the development of these circuits in fully monolithic integrated circuit technology, where all or most of the required circuit components are realized on a single chip of silicon. Offering various output current and voltage ratings, and output voltages of either positive or negative polarity, several commercial regulator integrated circuits are now available to suit the requirements of most applications. The designs of these regulators have matured and have become rather sophisticated. In addition to implementation of the high-gain feedback amplifier, the series pass element, and an accurate voltage reference, all on a single silicon die, built-in protection against overload conditions (such as output short circuits and excessive operating temperature) is now standard. Novel circuit-design, processing, and packaging techniques have been developed and implemented to achieve increased accuracy, temperature stability, efficiency, reliability, and power-handling capability, while reducing package size and cost. *See* INTEGRATED CIRCUITS.

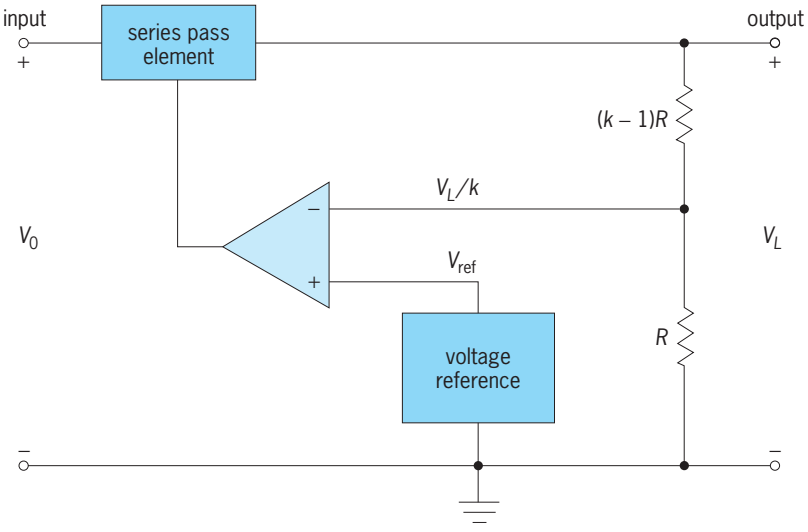


Fig. 2. Voltage regulator feedback circuit.

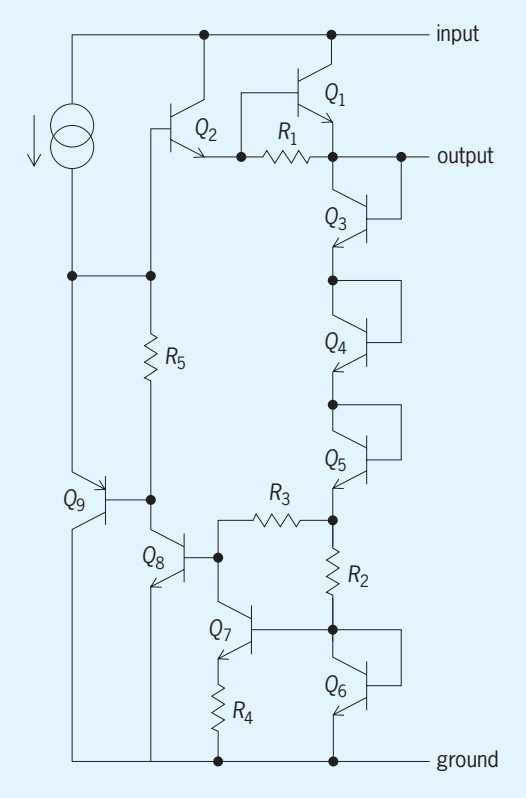


Fig. 3. Simplified circuit schematic of the LM109 voltage regulator. (*After R. J. Widlar, New developments in IC voltage regulators, IEEE J. Sol. State Circ., SC-6:2–7, February 1971*)

Example. In order to explain how monolithic voltage regulator chips work, a standard commercial integrated circuit is used as an example. A simplified schematic diagram showing the essential details of the regulator is given in **Fig. 3**. Although circuit design details vary considerably between commercially available regulators, depending on the design objectives, specifications, technology used, and manufacturer, the concepts implemented in the circuit of Fig. 3 are typical of the state of the art. This regulator integrated circuit was designed primarily for supplying power to logic circuits.

The circuit makes use of a basic property of bipolar transistors: the base-emitter voltage V_{BE} of a transistor has a negative dependence on temperature (that is, V_{BE} decreases with increasing temperature), while the difference ΔV_{BE} in the base-emitter voltages of two transistors biased at constant currents is directly proportional to temperature. Thus, it is possible to generate a weighted sum of V_{BE} and ΔV_{BE} of the form of Eq. (3), and to choose the weights k_1 and k_2 such

$$V_{\text{ref}} = k_1 V_{BE} + k_2 \Delta V_{BE} \tag{3}$$

that the two temperature variations cancel, resulting in a reference voltage V_{ref} that is insensitive to temperature.

Referring to the circuit shown in Fig. 3, the voltage across resistor R_4 is given by $V_{BE6} - V_{BE7} = \Delta V_{BE}$. Since transistor Q_7 and resistor R_3 conduct approximately the same current as R_4, the voltage across R_3 is $(\Delta V_{BE}/R_4)R_3$. The output voltage is simply the sum of the base-emitter voltages of transistors Q_3–Q_5, the voltage across R_3, and the base-emitter voltage of Q_8. Assuming that Q_3–Q_5 and Q_8 have the same base-emitter voltage V_{BE}, the output voltage can be written as Eq. (4), which has the same form as Eq. (3)

$$V_{\text{ref}} = \frac{R_3}{R_4} \Delta V_{BE} + 4V_{BE} \tag{4}$$

with $k_1 = (R_3/R_4)$ and $k_2 = 4$. Thus, by appropriately choosing the resistance ratio R_3/R_4, a temperature-stable reference is obtained.

Comparing this circuit with the block diagram of Fig. 2, transistors Q_8 and Q_9 perform the amplifying function shown, while Q_1 is the series pass element that supplies the load current. An additional device Q_2 is used to provide the required base-current drive for Q_1. In this circuit, the output load voltage is the samee as the generated reference voltage [that is, $k = 1$ in Eq. (2)].

The complete regulator circuit is based essentially on the simplified schematic of Fig. 3, but includes additional circuitry for improving accuracy and protecting the device under adverse operating conditions. For example, the output current is sensed on-chip and, in case of an accidental short circuit in the load, this current is limited to a safe value. However, even though the output current is limited, excessive power dissipation can cause the chip to overheat. Additional thermal protection circuitry included on-chip further limits the output current when the chip temperature reaches about 175°C (347°F).

Characteristics. This monolithic regulator provides a fixed 5-V output voltage at load currents up to 1 A and is extremely simple to use. The only external components required are two small capacitors (connected from the input and output terminals to ground) for improving high-frequency stability and transient response. The output voltage of the regulator changes by less than 50 mV for input voltages ranging from 7 to 35 V and load currents from 0 to 1 A. The temperature drift in the output voltage is less than 0.02%/°C (0.01%/°F) over the rated ambient operating temperature range (−55°C to 125°C or −67 to 257°F).

Ashok P. Nedungadi

Power-System Voltage Regulators

Voltage regulators are used in the generation, transmission, and distribution of electric power.

Power transmission and distribution. Voltage regulators are used on distribution feeders to maintain voltage constant, irrespective of changes in either load current or supply voltage. Voltage variations must be minimized for the efficient operation of industrial equipment and for the satisfactory functioning of domestic appliances, television in particular. Voltage is controlled at the system generators, but this alone is inadequate because each generator supplies many feeders of diverse impedance and load characteristics. Regulators are applied either in substations to control voltage on a bus or individual feeder or on the line to reregulate the outlying portions of the system. These regulators are variable autotransformers with the primary connected across the line. The secondary, in which an adjustable voltage is induced, is connected in series with the line to boost or buck the voltage. *See* AUTOTRANSFORMER; ELECTRIC DISTRIBUTION SYSTEMS; ELECTRIC POWER SUBSTATION.

A control and drive provide automatic operation. A voltage-regulating relay senses output voltage. When the voltage is either above or below the band of acceptable voltage maintained by the regulator, this relay causes the motor to operate and change the regulator position to raise or lower the voltage as required to bring it back within the control band. It is desirable to maintain constant voltage at the average load center out on the feeder rather than at the regulator terminals. Hence the control circuit includes a line-drop compensator with resistance and reactance elements that can be adjusted to represent line impedance. These impedances carry current proportional to circuit load, thereby simulating the voltage drop between the regulator and the load center, and modify the voltage sensed by the voltage-regulating relay. *See* RELAY.

Two principal types of feeder regulators are used: the step regulator, which provides increments or steps of voltage change, and the induction regulator, which provides continuous voltage adjustment.

Step voltage regulator. The transformation ratio of the autotransformer in this regulator is adjusted by a voltage selector switch which changes the secondary winding tap connected to the line. **Figure 4** shows the most commonly used circuit.

Switching is performed without interrupting load current by means of the two switch fingers in the selector switch. When the switch moves from the full-tap position shown in Fig. 4, one finger contacts the next tap before the other finger leaves the first tap; this constitutes a tap-bridging position. Switching reactors limit the current circulating between the bridged taps. Most regulators are designed to operate continuously in these bridging positions (as well as the full-tap positions) to provide voltages midway between the voltages of adjacent taps. Thus the voltage

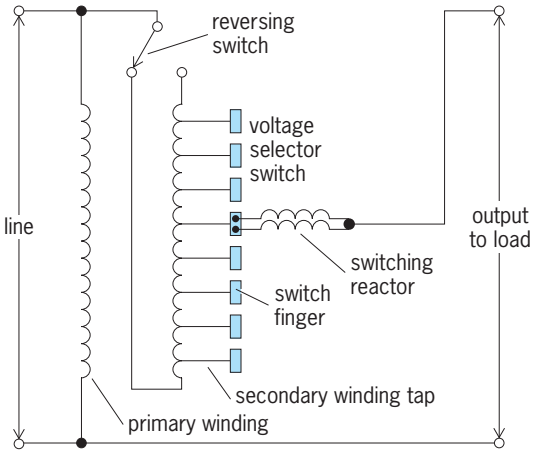


Fig. 4. Step voltage regulator circuit.

step between adjacent positions is half-tap voltage, and the number of winding taps required is half the number of operating positions in the boosting range. The automatic reversing switch changes the polarity of the secondary relative to the primary, thus providing a bucking range equal to the boosting range. Single- and three-phase designs are available with a range of ±10% in 16 or 32 steps. Other design variations may employ single-finger switching and operate continuously only in full-tap positions.

Load tap-changing transformers are often used to provide both regulation and transformation from one voltage level to another. They are similar to step voltage regulators, except that they have separate high- and low-voltage windings.

Induction voltage regulator. This is similar in structure to a wound-rotor induction motor with rotor restrained so that it moves only to adjust voltage. The primary winding on the rotor is magnetically coupled with the series secondary winding on the stator. *See* INDUCTION MOTOR; MOTOR.

The principle of a two-pole single-phase regulator is illustrated in **Fig. 5**. Secondary voltage is continuously adjusted from full buck to full boost by changing the relative angular position of these windings through 180 electrical degrees.

Single-phase regulators require an additional permanently short-circuited rotor winding that is in space quadrature to the primary winding. Without this winding, the reactance of the regulator to the line current flowing in the secondary would be excessive in the neutral region between buck and boost. Three-phase induction regulators, if built on a three-phase core, do not require a short-circuited winding. Such induction regulators inherently introduce phase shift between primary and secondary voltages and are no longer supplied for feeder regulation.

Other regulators. Other types of regulators are also used outside the United States. One construction is a transformer structure with moving coils to change coupling; another has contacts moving over the exposed conductors to provide a large number of small, discrete steps.

Line voltage may be increased by drawing leading-power-factor current through the line reactance. Static capacitors, shunt-connected in fixed or automatically switched banks, are often applied near the loads to raise voltage. The increase in voltage is not limited solely to the vicinity of the capacitor. They also help compensate for the usual system condition

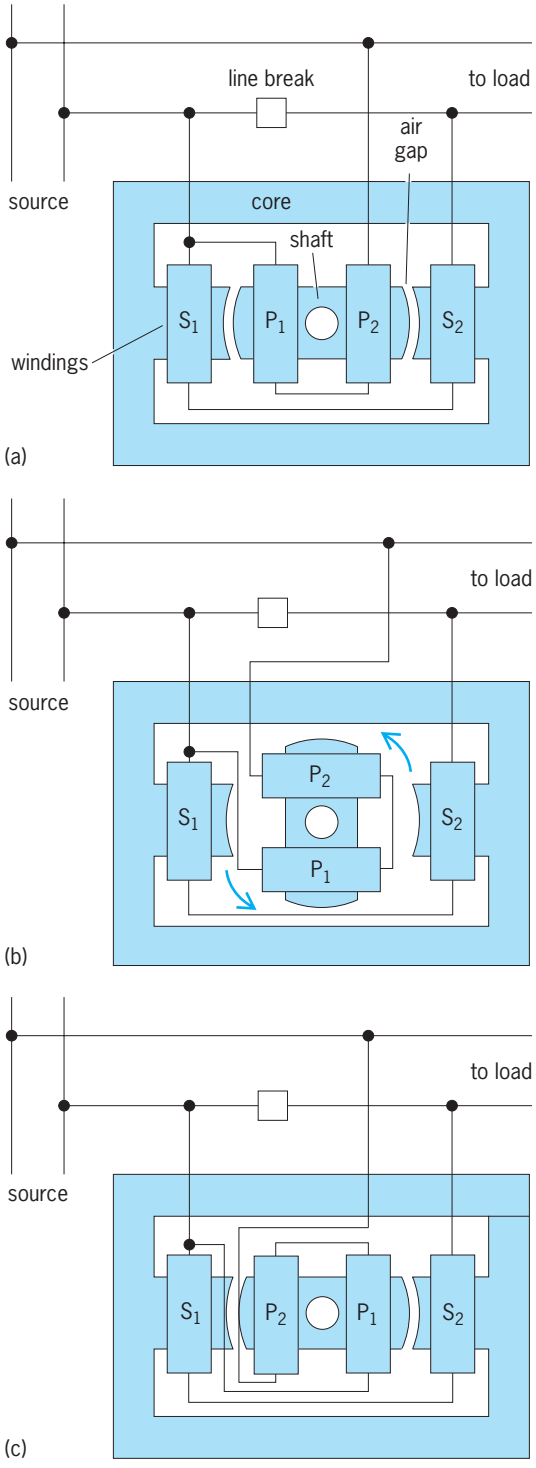


Fig. 5. Induction voltage regulator. (*a*) Full boost position. (*b*) Neutral position. (*c*) Full buck position of rotor. *P* coils are rotor primary coils, *S* coils are stationary secondary coils. (*After B. G. A. Skrotski, ed., Electric Transmission and Distribution, McGraw-Hill, 1954*)

of lagging power factor. The application of series capacitors on fluctuating loads is increasing.

Infrequently, synchronous machines called synchronous condensers, overexcited to draw leading current, are connected to lines. Their use has decreased with the availability of low-cost reliable static capacitors.

Conditions may exist where the inherent static capacity of the circuit is excessive. The line voltage may rise on cable systems or lightly loaded lines; shunt reactors may be used to neutralize the capacity of such systems. *See* STATIC VAR COMPENSATOR.

Donnell D. MacCarthy

Power generation. Voltage regulators are used on rotating machines in power generation applications to automatically control the field excitation so as to maintain a desired machine output voltage. Rotating machines, both small (down to 1 kW) and large (up to 1,000,000 kW), are the predominant means of power generation throughout the world, and voltage regulators of varying design and sophistication are employed on most of them. Even ac generators (or alternators) in automotive applications employ voltage regulators utilizing similar principles. *See* ALTERNATING-CURRENT GENERATOR; DIRECT-CURRENT GENERATOR; ELECTRIC POWER GENERATION; ELECTRIC ROTATING MACHINERY; GENERATOR.

The basic concept of a voltage regulator applied to an ac machine is shown in **Fig. 6**. The voltage regulator senses the output voltage of the ac machine (either directly in a low-voltage machine or through step-down voltage transformers in high-voltage machines) and compares this voltage feedback signal with a voltage reference. This voltage reference can be either fixed or adjustable. The difference between the reference (desired) and feedback (actual) voltages is the voltage error signal. The voltage regulator amplifies this error signal and in more sophisticated systems provides various stabilizing and limiting functions. The resulting voltage regulator output will determine the ac generator field excitation, either directly or through an exciter.

When the voltage error signal is positive, the voltage regulator will act to boost (increase) excitation to the field winding to help maintain the desired output voltage. When the error signal is negative, the voltage regulator will act to buck (decrease) excitation. In most generator applications it is not feasible to operate with fixed, uncontrolled field excitation, because the required level of field excitation to maintain rated output voltage increases as active power [measured in kilowatts (kW)] or reactive power [measured in kilovolt-amperes reactive] (kVAr)] loading is increased. Without a voltage regulator, the generator voltage will typically drop to unacceptable levels as loading is increased, and potentially will cause loss of synchronism with the power system for applications in which the generator is synchronized to a power grid. The voltage regulator will sense the tendency of the generator output voltage to drop with increasing loading, and will appropriately increase field excitation level in order to maintain the desired output voltage automatically.

Exciters. In most utility power generation applications, the ac generator is a synchronous machine, having a rotating dc field winding and a stationary three-phase armature winding. In these cases, the voltage regulator senses the output or terminal voltage of the three-phase armature winding and adjusts the excitation level to the field winding. In large synchronous machine applications, the power requirements of the field winding are significant enough to require a device called an exciter. The exciter can be either a static power amplifier comprising solid-state power conversion devices (such as thyristors or silicon controlled rectifiers) or a rotating machine, which can be either a dc machine (in older applications) or an ac machine whose output is rectified. The dc output of a static or rotating exciter is then delivered to the rotating synchronous machine field winding by means of slip rings. A brushless exciter is a type of rotating exciter that has a rotating armature and rectifier and thus provides field excitation without the need for slip rings. The combination of the voltage regulator and exciter is commonly called the

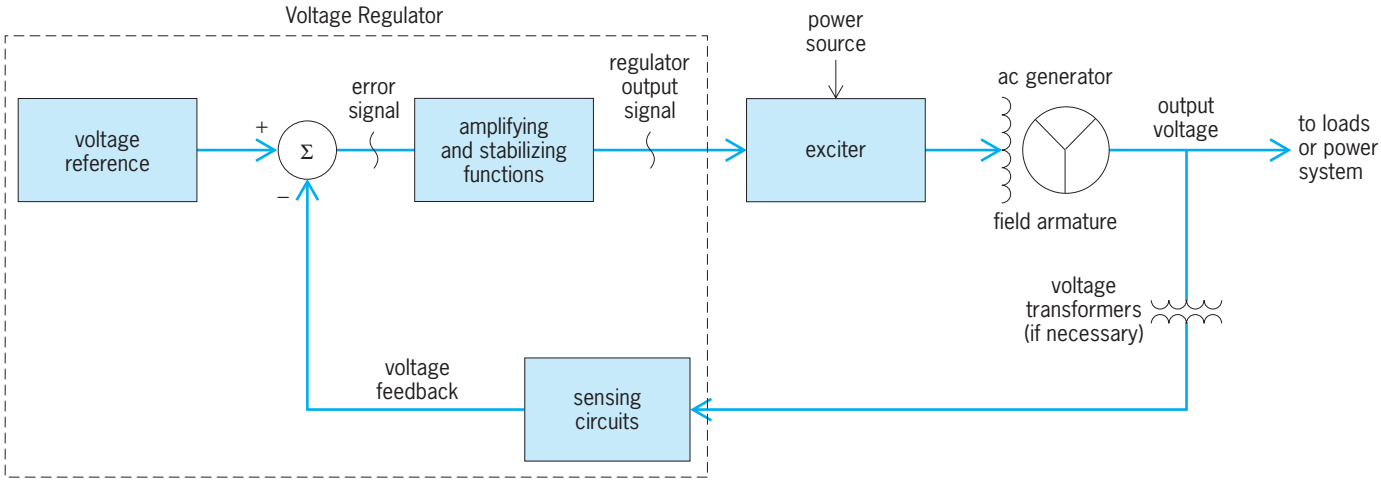


Fig. 6. Simplified block diagram of voltage regulator application.

excitation system. *See* SEMICONDUCTOR RECTIFIER; SLIP RINGS.

Regulator technology. Early voltage regulators were electromechanical devices which controlled exciter output by turning large rheostats or switching resistors with movable contacts. These regulators were slow and often required a deadband to prevent moving parts from quickly wearing out. The first solid-state continuously acting voltage regulators employed magnetic amplifiers for control. Modern analog voltage regulators utilize integrated circuits containing operational amplifiers and other electronic devices. Digital microprocessor or computer-based voltage regulators are becoming more prevalent. *See* MICROPROCESSOR; OPERATIONAL AMPLIFIER; RHEOSTAT.

Forcing capability. In addition to simply regulating steady-state generator output voltage, voltage regulators as part of an excitation system typically provide transient forcing capability during short circuits and other disturbances, allowing the exciter output to temporarily increase above that required for steady-state conditions. This capability helps the generator to more quickly recover from such faults, and thus enhances its stability or its tendency to remain in synchronism with the power system.

Limiting and portective functions. Modern voltage regulators also usually provide various limiting the protective functions which are designed to protect both the excitation system and the generator from such conditions as overexcitation (which would overload the exciter or overheat the generator field winding), underexcitation (which would jeopardize the stability of the generator connection with the power system), generator overvoltage, and excessive generator flux (as sensed by the ratio of generator output voltage versus frequency). The limiting functions provide control action to override the normal voltage regulator control signal in an attempt to provide continued operation while operating within the capability of the equipment. The protective functions back up the limiting functions and automatically remove the generator from service if the limiting functions fail to keep the operating quantities within limits.

J. D. Hurley

Bibliography. P. R. Gray and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, 4th ed., Wiley, 2001; *National Semiconductor Linear Data Book*, vol. 1, 1987; R. J. Traister, *Voltage Regulator Circuit Manual*, 1989.

Volume control systems

Systems that maintain proper audio signal levels in applications such as sound recording, public address systems, and broadcasting. Two types of electronic devices, compressors and limiters, perform this operation automatically without the need for human intervention, but differ in the way that they perform.

A compressor slowly varies its gain so as to maintain its output volume level at some constant average value. When used to process the audio signal from a microphone, a compressor equalizes the loudness of speech from different talkers. It compresses the dynamic range of the signal. A limiter controls the peak levels of an audio signal. Limiters are used to prevent overmodulation of transmitters in broadcast facilities, to prevent peak clipping in public address system audio amplifiers, and to prevent overload of audio recorders. A compressor followed by a limiter can be used to dramatically enhance the loudness of an audio signal without an increase in the peak level. This is a common practice in broadcasting to increase signal range and to maximize program loudness. *See* GAIN.

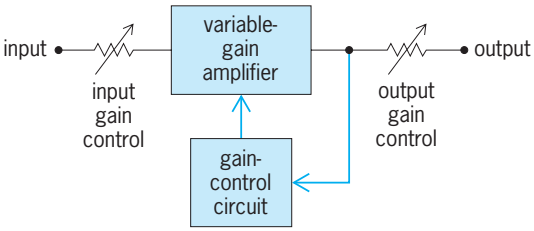


Fig. 1. Basic block diagram of a compressor and a limiter.

The input of a compressor or a limiter is applied to a variable-gain amplifier (**Fig. 1**). A gain-control circuit monitors the level of the output and generates a voltage which controls the amplifier gain. Manual controls at the input and output allow setting the levels for proper interface to external devices. *See* AUTOMATIC GAIN CONTROL (AGC).

Figure 2 illustrates the typical variation of output level versus input level for a compressor. The threshold V_t is the input voltage below which the

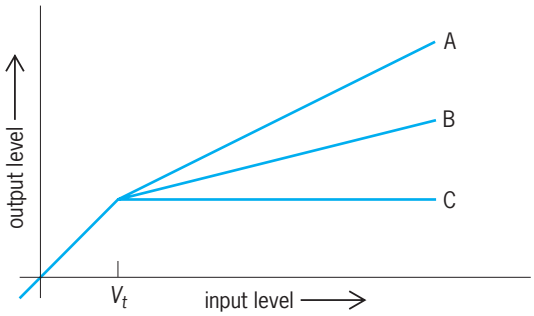


Fig. 2. Variation of compressor output level versus input level for three compression ratios.

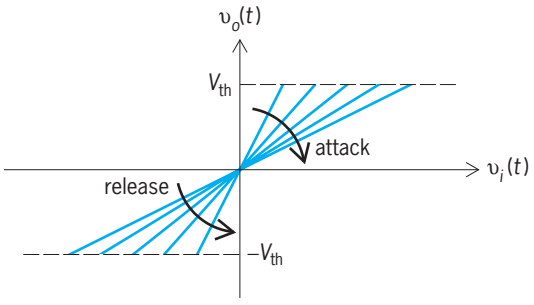


Fig. 3. Dynamic output-versus-input [$v_o(t)$-versus-$v_i(t)$] characteristic of a limiter, showing how the limiter gain decreases when its peak output voltage "tries" to exceed the threshold voltage V_{th}.

compressor gain does not vary with input. For an input greater than V_t, curves are shown for three different compression ratios. The compression ratio is the decibel change in input level required for a 1-dB change in output level. For curve A, it is 2 dB; for curve B, it is 4 dB; for curve C, it is infinite. When the input level increases, the time required for the gain to decrease is called the attack time. When the input level decreases, the time required for the gain to increase is called the release time. Typically, the attack time for a compressor is 10–100 ms and the release time is 1 to 10 s. A gated compressor is one in which the gain "freezes" when the input signal falls below threshold. This prevents low-level background sounds from being increased in amplitude.

Figure 3 illustrates the dynamic output-versus-input [$v_o(t)$-versus-$v_i(t)$] characteristics of a limiter. The gain corresponds to the slope of the $v_o(t)$-versus-$v_i(t)$ curve. For an output voltage in the range $-V_{th} \leq v_o(t) \leq V_{th}$, the gain is constant. When $v_o(t)$ "tries" to leave this range, the gain drops rapidly to limit the peak output level. This is indicated by the clockwise rotation of the slope on the graph. After the peak, the gain gradually returns to its original value. Typically, a limiter has an attack time of 1–10 ms and a release time of 10–100 ms. *See* LIMITER CIRCUIT; RADIO BROADCASTING; SOUND RECORDING; SOUND-REINFORCEMENT SYSTEM; SOUND-REPRODUCING SYSTEMS.

W. Marshall Leach, Jr.

Bibliography. G. Davis and R. Jones, *Sound Reinforcement Handbook*, 2d ed., 1990; J. M. Eargle, *Handbook of Recording Engineering*, 3d ed., 1996; D. M. Huber and R. E. Runstein, *Modern Recording Techniques*, 5th ed., 2001; D. Lyver and G. Swain son, *Basics of Video Sound*, 2d ed., 1999.

Volume unit (vu)

A unit used to measure the strength of electrical waves produced by microphones. In sound recording, broadcasting, and public address systems, microphones convert acoustical signals into electrical waves that are nonperiodic and cannot be simply described in terms of voltage, current, power, or frequency. To provide a practical means of assigning a numerical value to the strength of such waves, the concept of "volume" is used. The volume of an audio program wave is the magnitude of the wave as measured with a standard volume indicator. This is a meter calibrated to read in volume units (vu, pronounced vee-you).

A vu meter consists of an attenuator, a rectifier, and a readout device. The latter is usually an electromechanical d'Arsonval meter or a light-emitting diode (LED) array. In **Fig. 1** the block diagram of a vu meter is shown with a sine-wave input. The attenuator output is applied to the full-wave rectifier which drives the meter. The meter reading is proportional to the amplitude of the input. **Figure 2***a* illustrates a vu meter with a d'Arsonval meter readout. The level is indicated by the motion of the meter needle. Figure 2*b* illustrates a vu meter with an LED readout. The level is indicated by the cell that is lit from behind by an LED. Unlike the d'Arsonval meter, the LED array must be driven by an electronic circuit. *See* LIGHT-EMITTING DIODE; RECTIFIER.

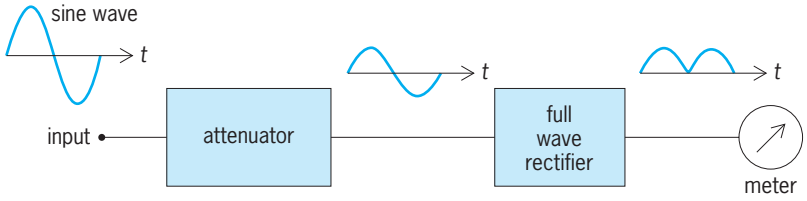


Fig. 1. Block diagram of a vu meter showing voltage waveforms for a sine-wave input.

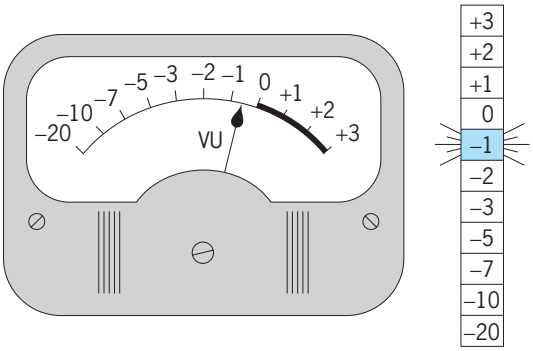


Fig. 2. Typical readouts for (*a*) a d'Arsonval vu meter and (*b*) an LED array vu meter.

The attenuator is adjusted so that the meter reads 0 vu with a sine-wave input of 1 kHz and an amplitude that depends on the application. For example, it might be set so that 0 vu represents 1 V rms. With the sine-wave input, the meter is calibrated so that its reading corresponds to the decibel (dB) change in voltage from the reference level. *See* DECIBEL; SINE WAVE.

William M. Leach, Jr.

Bibliography. G. Davis and R. Jones, *Sound Reinforcement Handbook*, 1990; J. M. Eargle, *Handbook of Recording Engineering*, 1996; D. M. Huber and R. E. Runstein, *Modern Recording Techniques*, 1995; D. Lyver, *Basics of Video Sound*, 1995.

Volumetric efficiency

In describing an engine or gas compressor, the ratio of volume of working substance actually admitted, measured at specified temperature and pressure, to the full piston displacement volume. For a liquid-fuel engine, such as a diesel engine, volumetric efficiency is the ratio of volume of air drawn into a cylinder to the piston displacement. For a gas-fuel engine, such as a gasoline engine with carburetor, throttle body, or port injection, volumetric efficiency is based on the charge of fuel and air drawn into the cylinder. *See* COMPRESSOR; ENGINE.

Volumetric efficiency of naturally aspirated automobile and aircraft reciprocating engines may be 85–90% at rated speed. Supercharging or turbocharging increases volumetric efficiency, giving values

over 100%. At low speeds and wide-open throttle, volumetric efficiency is high. At high speeds, resistance to airflow through the carburetor or fuel-injection throttle body, the manifolds, and valve ports decreases volumetric efficiency. Two-cycle engines with crankcase compression may have a volumetric efficiency of about 60% at low speed and less than 50% at high speed. Air compressors, refrigerator compressors, and dry vacuum pumps are generally specified for a volumetric efficiency of 60–90%, dependent upon the compressor ratio, type of valve, valve gear, and machine speed. *See* INTERNAL COMBUSTION ENGINE; RECIPROCATING AIRCRAFT ENGINE.

Donald L. Anglin

Bibliography. W. H. Crouse and D. L. Anglin, *Automotive Engines*, 8th ed., 1994.

Volvocales

A large order of green algae (Chlorophyceae) comprising all forms that normally are flagellate and motile. In zoological classification, it is called Volvocida and placed in the class Phytomastigophora. The cells are solitary or united into colonies of definite structure (coenobia), often with morphological and functional differentiation among the component cells. Some taxonomists place the unicellular forms in a separate order, Chlamydomonadales. Unicells that have volvocalean cytological features but are nonflagellate and sedentary in their vegetative phase are considered here to constitute the order Tetrasporales. Alternatively, these sedentary forms may be retained in the Chlamydomonadales or Volvocales. *See* CHLOROPHYCEAE.

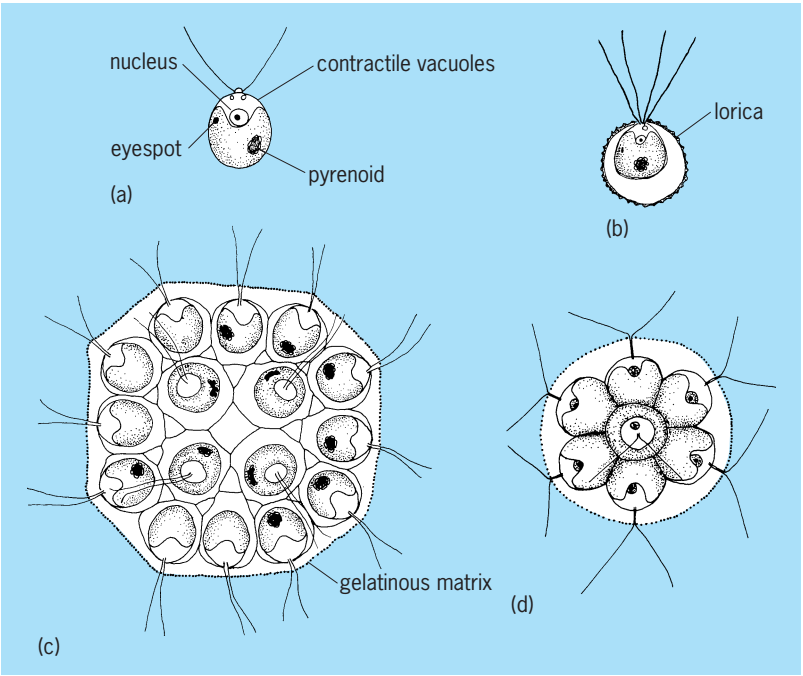


Fig. 1. Representative genera of Volvocales. (*a*) *Chlamydomonas* (Chlamydomonadaceae). (*b*) *Pedinopera* (Phacotaceae), showing granulate lorica in optical section. (*c*) *Gonium* and (*d*) *Pandorina*, both members of the Volvocaceae.

Unicellular forms. *Chlamydomona* (**Fig. 1***a*), which is presumed to be similar to ancestral Volvocales, has a very small cell body (usually 7–40 micrometers in greatest dimension) that is spherical, ellipsoid, or pyriform. The wall (or theca), which consists primarily of glycoproteins, is usually smooth, but may have blunt protuberances. The cell bears two smooth flagella of equal length at its apex, with contractile vacuoles and an eyespot at the anterior end. The chloroplast is usually cup-shaped, but it may be H-shaped, perforate, or stellate. Chloroplasts contain one or more pyrenoids. As in all Volvocales, the cells are uninucleate.

Other genera with chlamydomonad features are distinguished on the basis of flagellar number, presence or absence of photosynthetic pigments, presence or absence of pyrenoids, and cell shape. Together with *Chlamydomonas* they constitute the family Chlamydomonadaceae. Two other families of unicellular Volvocales are recognized: Dunaliellaceae, in which the unicells are naked; and Phacotaceae, in which the unicells are surrounded by a lorica—a rigid, hyaline or dark brown, wall-like structure, which is often impregnated with compounds of calcium, iron, or manganese and which may be sculptured (Fig. 1*b*).

Asexual reproduction in unicellular Volvocales is by bipartition (frequently repetitive) of a protoplast, often during a transitory nonmotile period. The progeny is released by rupture or gelatinization of the parental wall. Sexual reproduction is isogamous, anisogamous in varying degrees, or oogamous. The zygote develops a thick ornamented wall that protects it during adverse periods (as when a pool becomes dry). Meiosis occurs during germination of the zygote, so that all stages in the life history except the zygote are haploid. Many species have a temporary nonmotile gelatinous stage (palmella stage) in which the cells readily develop flagella and again become motile.

Unicellular Volvocales are ubiquitous in fresh, brackish, and coastal marine waters and in such terrestrial habitats as soil, snow, and ice. They are especially abundant in organically enriched bodies of fresh water. Nearly 50 genera with more than 800 species are recognized. Because it is common and easily cultured, *Chlamydomonas* is frequently used as an experimental organism in cytological, biochemical, and genetic studies. Certain physiological strains of *Dunaliella* are grown in mass culture as a source of carotenoids.

Colonial (coenobial) forms. Most colonial Volvocales are placed in the family Volvocaceae. In this family, coenobia are composed of *Chlamydomonas*-like biflagellate cells embedded in a firm gelatinous matrix. Coenobia have the form of a slightly curved plate or are ellipsoid to spherical. The component cells are arranged according to a predetermined pattern. A phylogenetic series, involving an increase in number of cells and the development of polarity, results in morphological and functional differentiation of the component cells. Their number is a multiple of 2, ranging from 4 (*Gonium*) to about

50,000 (*Volvox*). Each cell is surrounded by a gelatinous envelope, which may remain distinct or become confluent with those of adjacent cells. Cytoplasmic strands connect adjacent cells in *Gonium* and in some species of *Volvox*. Vegetative cells are spherical, ovoid, pyriform, or hemispherical.

Asexual reproduction is by repeated bipartition of the protoplast of an ordinary or specialized cell to form a new coenobium (autocoenobium) that is a miniature of the parent. All cell divisions are longitudinal and in a regular sequence. At the eight-celled stage, the autocoenobium has the form of a curved plate. In more advanced genera, further divisions produce a hollow sphere of very small cells, which sinks into the gelatinous matrix of the parent. There is an opening to the surface at the anterior end of the developing autocoenobium (the phialopore). The flagella are exserted into the cavity. After the last division, the phialopore expands, its lips fold back, and the hollow sphere turns inside out (everts) so that the flagella regain their normal orientation, directed outward from the coenobium. The daughter coenobium revolves slowly within the enlarged gelatinous wall of the parental cell and eventually escapes through a pore to the outside.

Sexual reproduction forms a phylogenetic series from isogamy through anisogamy to oogamy. The zygote secretes a thick wall and, after a period of dormancy, germinates meiotically to produce one to four zoospores which, like motile gametes, are biflagellate. Each zoospore divides repeatedly to form a coenobium (except in *Gonium*, where the four zoospores remain together as a coenobium, even though they are genetically heterotypic).

Gonium (Fig. 1*c*), with 4–32 cells arranged in a slightly curved quadrangular plate, is the simplest member of the family. All cells form autocoenobia simultaneously. Sexual reproduction is isogamous. In *Pandorina* (Fig. 1*d*), 16 (rarely 8 or 32) cells are arranged compactly in a hollow subspherical or ellipsoid coenobium. All eyespots are the same size, or those on one side are markedly larger, thus establishing polarity. All cells form autocoenobia simultaneously. After the last division, the bowl-shaped juvenile coenobium everts to form a sphere. Sexual reproduction is anisogamous. In *Volvulina*, the cells (usually 16) are hemispherical, the flat side facing outward, and are arranged in transverse tiers at a conspicuous distance from one another. Individual cell envelopes are visible. Eyespots in the anterior tier are larger than those in other tiers and may be lacking in the posterior half.

The coenobium of *Eudorina* comprises 8–64 (usually 32) spherical cells embedded in a homogeneous spherical or obovoid matrix that may have mammillate projections posteriorly. The cells are often in distinct transverse tiers (4–8–8–8–4). In certain species the eyespot is progressively smaller from anterior to posterior poles and may be lacking in the lowermost tier. All cells of a coenobium, or less often only certain cells, produce autocoenobia asexually in a manner typical of the family. Sexual reproduction involves an advanced type of anisogamy

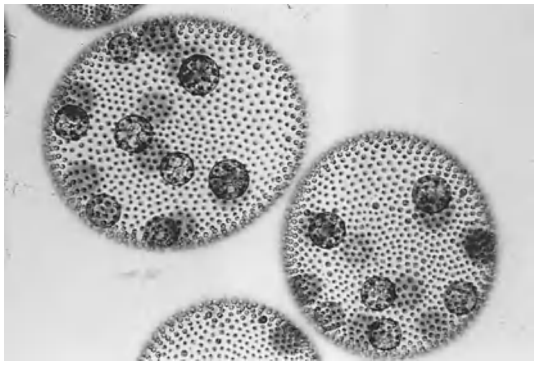


Fig. 2. ***Volvox carteri*, young coenobia showing large gonidia. (*R. C. Starr, University of Texas*)**

in which a packet of spindle-shaped male gametes resulting from repeated bipartition of a cell of a male coenobium swims as a unit to the female coenobium, where it dissociates into individual gametes. Although the female gamete is flagellate, it retains its position within its coenobium during sexual fusion.

Platydorina has flattened, horseshoe-shaped coenobia, with the cells (32, 64, or 128) arranged in one layer. Depending upon the species, certain cells, varying from 4 to half the coenobium, remain small and vegetative. These may be at the anterior pole or scattered among the remaining cells, which are two or three times larger. The large cells (gonidia), which often lose their eyespot and flagella, have a massive chloroplast with several pyrenoids. All gonidia may divide simultaneously to produce autocoenobia. Sexual reproduction is similar to that in *Eudorina*, except that the female gamete may lose its flagella prior to fertilization.

Coenobia of *Volvox* (**Fig. 2**) are relatively large (0.5–1.5 mm in diameter), comprising about 500 to 50,000 cells in a single layer just within the periphery of a spherical or ellipsoid gelatinous matrix. Individual gelatinous cell envelopes are visible in most species. The cavity is filled with a watery gelatinous substance. In all embryonic coenobia, cells have cytoplasmic connections. During maturation these may either break or remain as delicate or broad strands. Eyespots decrease in size from anterior to posterior poles. During maturation of a coenobium, certain cells (2–50) in the posterior half differentiate into gonidia by enlarging, producing additional pyrenoids, and losing their eyespot and flagella. The other, smaller cells of the coenobium remain vegetative. Development of a gonidium into a daughter coenobium is typical of the order, involving eversion (**Fig. 3***a*). Sexual reproduction is oogamous. Most species are heterothallic. Homothallic species produce either unisexual or bisexual coenobia. Male gametes are produced in a packet (curved plate or compressed sphere) by enlarged cells resembling gonidia in a manner similar to the asexual development of an autocoenobium (Fig. 3*b*). Cells that function as eggs also resemble gonidia (Fig. 3*c*). The packet of male gametes dissociates as it approaches an egg.

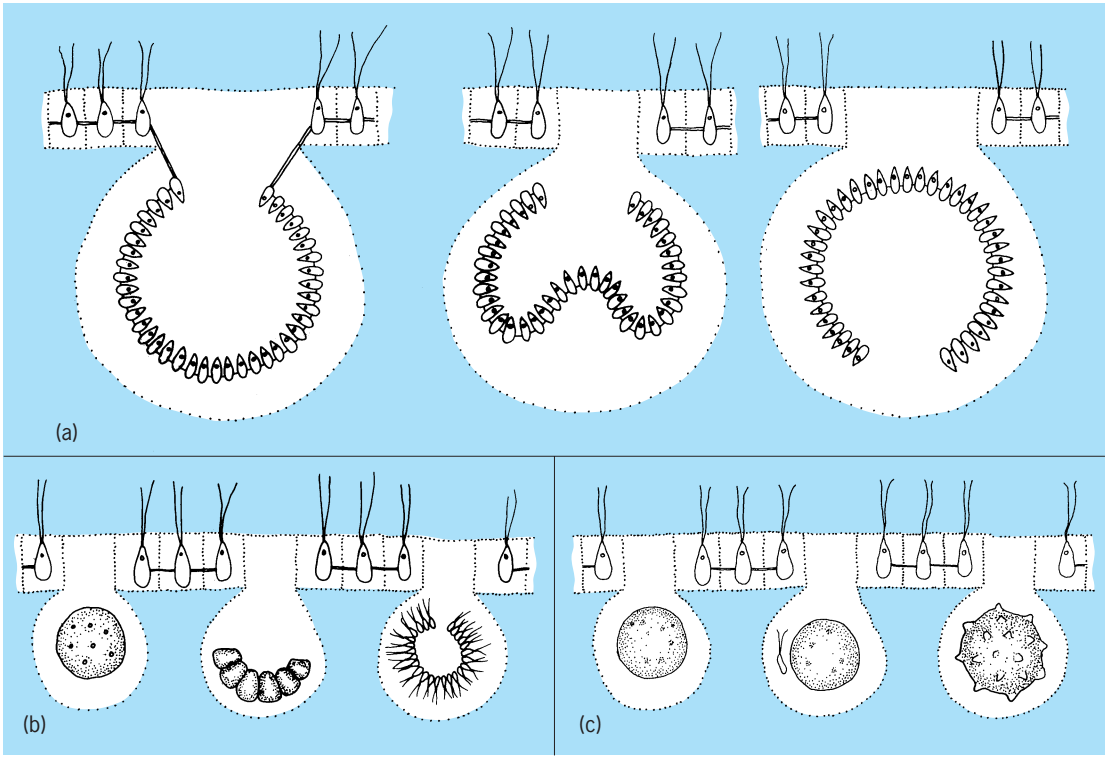


Fig. 3. Reproduction in *Volvox*. (*a*) Asexual development of an autocoenobium, showing eversion. (*b*) Development of packet of sperm from gonidium. (*c*) Fertilization of egg by sperm, resulting in spiny-walled zygote. (*After G. M. Smith, Cryptogamic Botany, 2d ed., McGraw-Hill, 1955*)

The zygote secretes a thick wall, accumulates hematochrome, and, following a period of dormancy, germinates meiotically. The result is a single zoospore, which divides repeatedly to produce a coenobium.

The Volvocaceae are abundant in organically enriched bodies of fresh water, especially temporary pools. There are 9 genera and about 48 species.

Paul C. Silva; Richard L. Moe

Bibliography. H. C. Bold and M. J. Wynne, *Introduction to the Algae: Structure and Reproduction*, 2d ed., 1997; G. M. Smith, *The Fresh-Water Algae of the United States*, 1950.

Volvocida

An order of the class Phytamastigophorea. The protozoa, also known as the Phytomonadida, are grass-green, but a few are colorless (*Polytoma*). Individual cells may be as small as 8 micrometers. They closely relate the flagellates to the algae. They have one flagellum (*Pedimonas*), two (*Chlamydomonas*), four (*Carteria*) (**Fig. 1**), or eight (*Polyblepharides*). Plural flagella are usually equal. The group is large and about one-fourth of the approximately 100 genera form palmelloid or dendroid colonies (*Tetraspora*, *Chlorangium*), with flagella only in reproductive cells. Cell walls are of cellulose and often they are thick. Chromatophores contain the same chlorophylls as higher plants. Pyrenoids have the usual form commonly found with starch as the reserve material. *See* CILIA AND FLAGELLA.

Unicellular species predominate, but at least 16 genera form colonies. These are irregular, linear, flat, or spherical. *Volvox globator* is a hollow sphere often 1 mm in diameter. Its mucilaginous periphery contains thousands of biflagellate zooids (**Fig. 2***a*). The zooids of *V. aureus* are pear-shaped and attached by threadlike protoplasmic strands to each other; the mucilage draws away from the thick outer boundary, but several cells are incorporated into a deep wedge

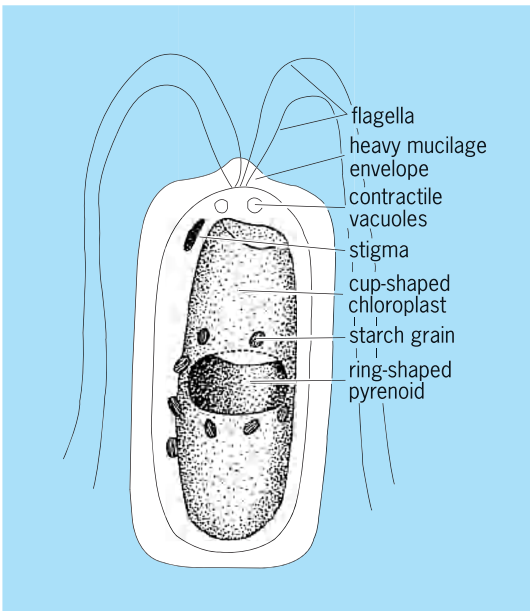


Fig. 1. *Carteria*, showing equal plural flagella.

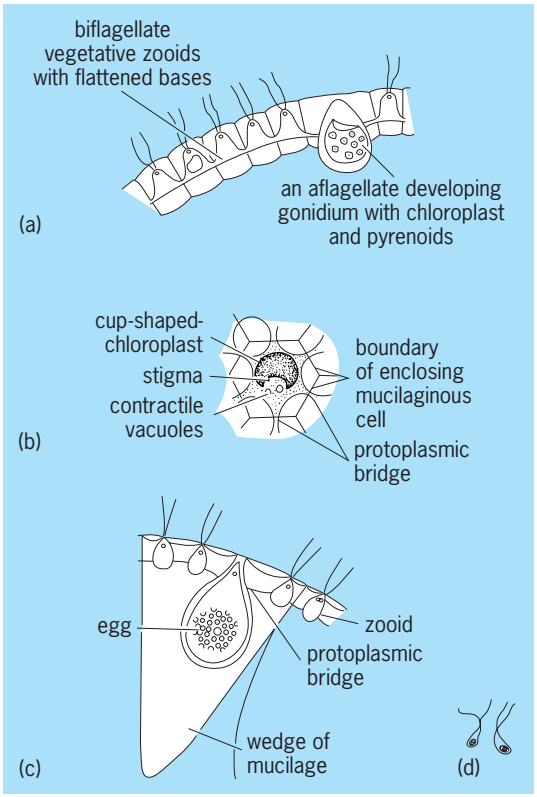


Fig. 2. Morphological features of *Volvox*. (*a*) *Volvox globator*, sectional view; (*b*) surface view of a single zooid. (*c*) *Volvox aureus*, sectional view; (*d*) sperm, enlarged about eight times by comparison with *c*.

of mucilage (Fig. 2*c*). Some zooids are gonidia that form new colonies asexually by a complicated multiple division. Isogamy, the production of like gametes which fuse, is common in Phytomonadida, but *V. globator* is monoecious, producing large eggs from certain zooids and packets of small sperm from others. Eggs, liberated into the hollow sphere, are fertilized by sperm. Thick rough walls envelop the fertilized eggs which eventually produce new colonies. Thus *V. globator* illustrates sexual reproduction comparable to that of higher animals. Most phytomonads are fresh-water inhabitants; some are terrestrial, some marine. *See* PHYTAMASTIGOPHOREA; PROTOZOA; REPRODUCTION (ANIMAL).

James B. Lackey

VOR (VHF omnidirectional range)

A short-range air navigation aid, which provides azimuth aid by visual means of cockpit instruments. A VOR system provides properly equipped aircraft with bearing information relative to the VOR station and magnetic north. The VOR system is used for landing, terminal, and en route guidance. It also gives virtually static-free regular weather broadcasts, special flight instructions, and voice and code station identification. The VOR service operates in the very high frequency (VHF) band between 108 and 118 MHz, sharing alternate channels with the localizer in the instrument landing system. Typically, VOR stations are co-located with a distance measuring equipment (DME) system or a tactical navigation (TACAN) system. The combined systems are referred to as VOR/DME or VORTAC stations and provide both azimuth and distance information. *See* DISTANCE-MEASURING EQUIPMENT; INSTRUMENT LANDING SYSTEM (ILS); TACAN.

VOR/DME beacons provide air navigation services throughout the world. However, the emergence of satellite navigation systems has motivated civil aviation administrations to reconsider their continued support of these services. In the United States, the Federal Radionavigation Plan of 2001 states that "VOR/DME will continue to provide navigation services for enroute through non-precision approach phases of flight throughout the transition to satellite-based navigation." The Federal Aviation Administration (FAA) plans to reduce (but not eliminate) VOR/DME services in the U.S. National Airspace beginning in 2010 based on the anticipated decrease in the use of VOR/DME for enroute navigation and instrument approach. Any reduction will occur in phases and VOR/DME will remain in service at selected airports where it will provide a backup to satellite-based services. *See* SATELLITE NAVIGATION SYSTEMS.

There are two types of VOR equipment. One employs a four-loop antenna array and is called conventional VOR. The other, Doppler VOR, having a 50-loop or larger antenna array located around a single-loop carrier antenna, is based on the Doppler principle and is designed for installation at locations which present especially difficult siting problems due to multipath conditions. Although the two types differ in design, from the standpoint of air navigation they function in essentially the same manner and can be received by the same equipment. Both types are used to define route intersections and most domestic airways. *See* DOPPLER EFFECT; DOPPLER VOR.

Operating principles. The VOR operates on the principle that the phase difference between two signals can be employed as a means of determining azimuth if one of the signals maintains a fixed phase through 360°, so it can be used as a reference, while the other is made to vary as a direct function of azimuth. The phase difference between these two signals will then equal the azimuth of the aircraft. In practice, two demodulated 30-Hz signals are used. These are called the reference-phase and variable-phase signals.

The standard antenna system for the conventional VOR transmitter is composed of four VHF loops, mounted in a square of 48 in. (1.22 m), above a metallic counterpoise. Each diagonal pair of loops is driven from the output of a goniometer (a mechanical beam-former) and produces a figure-of-eight sideband pattern which rotates clockwise. All four loops also are driven in-phase with carrier-frequency power to radiate a circular pattern. Addition of the two patterns causes a rotating limacon [a plane curve whose equation in polar coordinates (r, A) is $r = a \cos \theta + b$], which, when detected in an

airborne receiver, produces a variable-phase signal. *See* ANTENNA (ELECTROMAGNETISM).

A 9960-Hz subcarrier amplitude modulates the carrier frequency on the circular pattern. Furthermore, this subcarrier is frequency-modulated (FM) at 30 Hz with a deviation ratio of 16 for a total deviation of 480 Hz. The aircraft receiver separates the 30-Hz amplitude-modulated (AM) variable signal and the 9960-Hz subcarrier into two separate paths for processing. The 30-Hz signal is filtered and compared to the 30-Hz frequency-modulated signal from the subcarrier. The subcarrier is limited and demodulated by a frequency discriminator circuit. It is also filtered and compared to the 30-Hz amplitude-modulated signal for the azimuth selected by the pilot. *See* AMPLITUDE MODULATION; FREQUENCY MODULATION; FREQUENCY-MODULATION DETECTOR.

Most of the VOR stations within the United States include all-solid-state equipment and give remote indications of station availability to air-traffic control personnel and remote maintenance capability to technicians. This feature ensures reliability and reduces maintenance costs since one technician is able to service several stations.

Capabilities. The VOR system shows magnetic bearing to a VOR station and deviation from a selected course, in degrees. The ground-station errors are approximately $\pm 1.4^\circ$. A VOR system allows an essentially continuous update of deviation from a selected course based on internal operations at a 30-update-per-second rate. Initialization is typically less than 1 minute after turn-on. VOR systems have line-of-sight limitations that could limit ground coverage to 30 nautical miles (55 km) or less. At altitudes above 5000 ft (1500 m) the range is approximately 100 nmi (185 km), and above 20,000 ft (6000 m) the range will approach 200 nmi (370 km). Availability depends on the location of the VOR stations and is tied into these coverage considerations.

The capacity of VOR systems is unlimited. Their reliability approaches 100%, and no ambiguity is possible for a VOR system. VOR provides system integrity by removing a signal from use within 10 seconds of an out-of-tolerance condition detected by an independent monitor. *See* ELECTRONIC NAVIGATION SYSTEMS; RHO-THETA SYSTEM.

William I. Thompson III; Robert B. Flint; Richard L. Greenspan

Bibliography. P. Enge et al., Terrestrial radio navigation technologies, *Navigation* (Journal of the Institute of Navigation), 42(1):61–108, 1995; *1999 Federal Radionavigation Plan*, pp. C25–C27, U.S. Department of Defense Rep. DOD-4650.5 and U.S. Department of Transportation Rep. DOT-VNTSC-RSPA-98-1, Washington, DC.

Vortex

In common usage, a fluid motion dominated by rotation about an isolated curved line in space, as in a tornado, a whirlpool, a hurricane, or a similar natural phenomenon. The importance of vortices is due to two characteristics: general fluid flows can be represented by a superposition of vortices; and vortices, once created, have a persistence that increases as the effects of viscosity are reduced. The aerodynamic lift forces and most other contributors to the forces and moments on aircraft and other bodies moving through fluids do not exist in the absence of vortices. *See* AERODYNAMIC FORCE; HURRICANE; TORNADO; WATERSPOUT.

Vorticity. The strength of rotation is measured by a vector called the vorticity, ω, defined as the curl of the velocity vector. A flow devoid of vorticity is known as irrotational flow or potential flow, since the velocity vector can be expressed as the gradient of a scalar function, the velocity potential. The spatial distribution of the vorticity vector provides a precise characterization of the rotation effects in fluids, and the nature of what subjectively and popularly would be called a vortex. *See* CALCULUS OF VECTORS; LAPLACE'S IRROTATIONAL MOTION.

The vorticity vector field can be constructed by measuring the instantaneous angular velocity of small masses of fluid. The vorticity vector is twice the local angular velocity vector. Starting at any arbitrary point in the fluid, a line, called a vortex line, can be drawn everywhere parallel to the vorticity vector.

Vortex tube. A bundle of vortex lines defines a tubular region of space, called a vortex tube, with a boundary surface that no vortex line crosses.

Two simple rules follow from the definitions: (1) a vortex tube must either close on itself or end on a boundary of the fluid (including extending to "infinity" if the fluid is imagined to fill all space); and (2) at every cross section of a given vortex tube, the area integral of the normal vorticity has the same value at any given instant. The area integral is, by Stokes' theorem, equal to a line integral around the periphery of the tube, namely, the line integral of the velocity component parallel to the direction of the closed curve defining the line integral. This quantity is also known as the circulation around the line, so at an instant of time a vortex tube has a unique value of the circulation applying to all cross sections (**Fig. 1**). The circulation is a measure of the speed at which fluid circulates around the path.

In many flows of interest the magnitude of the vorticity vector is small except for the interior of one tube, or a relatively few of them. Each tubular surface defines a vortex core in which the vorticity vector is large. Such situations conform most closely to the popular description of a flow as a vortex. The

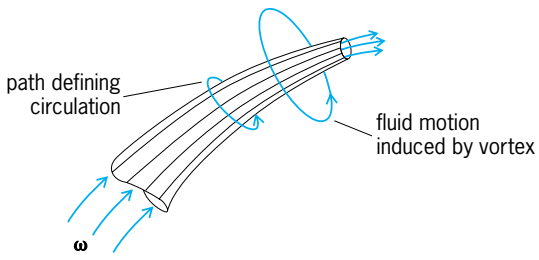


Fig. 1. Vortex tube; ω is the vorticity.

large angular velocity in the core causes a large centrifugal force on each element of fluid. To maintain mainly circular motion around the core, the centrifugal force must be balanced by a rapid decrease of pressure at the center of the vortex core. This large relative vacuum is responsible for most of the damage caused by concentrated atmospheric vortices such as tornadoes. The low pressure in the core is also responsible for the clear track of the wing-tip vortices trailing behind aircraft. These contrails are made visible by condensation of water vapor in the low-pressure cores. *See* CENTRIFUGAL FORCE.

Line vortex idealization. The core of a concentrated vortex usually has a diameter that is small compared to other dimensions associated with the vortex, such as its length or the radius of curvature of its center line. This leads to a convenient idealization, the line vortex or vortex filament, which takes the vorticity to be exactly zero outside the tube, and imagines the core area to be reduced to zero while keeping the circulation fixed. The model may be visualized by supposing that the vortex tube is straight and infinitely long, and all the motion is due to its presence. Then the motion in the irrotational region outside the core is said to be induced by the vortex, and consists of fluid moving in circular paths around the core with linear speed decreasing inversely with distance from the core center. While the fluid revolves around the vortex core, fluid particles in the irrotational region have no angular velocity around their mass centers; the situation is like a seat on a carnival ferris wheel. The resulting model has a singularity with infinite velocity at the vortex centerline—an unrealistic feature—but properly and conveniently represents the motion outside the core of the vortex being modeled.

Vortex sheet. Vortex lines confined to a layer rather than a tube describe fluid motion of a different character. This is most easily visualized when the direction of the vorticity does not vary, so all of the vortex lines are straight and parallel (**Fig. 2**). Assuming the vorticity has zero magnitude outside the layer, this vortex layer represents a flow with a different speed and direction on either side of the layer. Such a change in speed occurs at the edge of wakes produced by wind passing over an obstacle. Reducing the thickness of this layer of vorticity to zero leads to an idealization known as a vortex sheet, a surface in space across which there is a finite jump in velocity tangent to the surface (Fig. 2). Vortex sheets have a tendency to roll up because of self-induction. Aerodynamicists approximate flows in the wake of aircraft wings by vortex sheets, and the roll-up of this sheet leads to a pair of concentrated vortices trailing behind. The lift force on an aircraft depends on the formation of this trailing vortex system, and its strength is proportional to the weight of the aircraft. If the air is sufficiently humid, these vortices sometimes are visible due to condensation of water vapor in the vortex cores, which have low pressure and temperature. The trailing vortices in **Fig. 3** are made visible by injection of smoke near the wing tips. *See* WING.

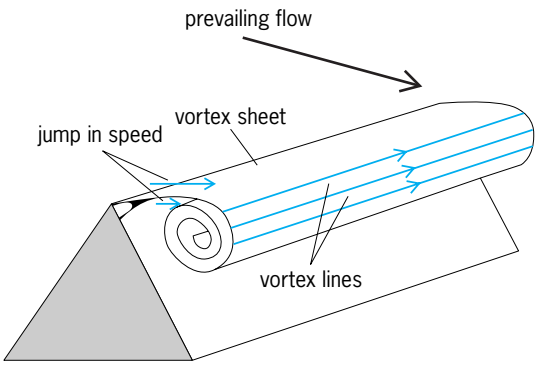


Fig. 2. Separation of flow from a sharp corner, showing a vortex sheet and its roll-up.

Fig. 3. Trailing vortex pair behind a Boeing 727 airplane (*NASA Dryden Flight Research Center file photograph ECN 3831*)

Persistence of vortex lines. In a fluid with spatially uniform density acted on by conservative external forces, the only mechanisms capable of changing the vorticity of a small element of fluid of fixed mass about its center of mass are frictional forces applied by the surrounding fluid by viscous action, and deformation of the element shape. If frictional effects are also sufficiently small to be ignored (a fluid under this approximation is called inviscid), then the circulation around any closed path that moves with the fluid is time independent. From this rule, known as Kelvin's theorem, it can be deduced that fluid particles with zero vorticity continue to have zero vorticity; and that fluid particles composing a vortex tube initially identified in a fluid continue to be a vortex tube, and the circulation for the tube is time independent. This result is usually paraphrased as "vortex lines move with the fluid." This conclusion can be extended to a special but important class of compressible motion known as barotropic flows, for

Fig. 4. Photograph of the Union City, Oklahoma, tornado of May 24, 1973. (*Courtesy of NOAA Photo Library, NOAA Central Library; OAR/ERL/National Severe Storms Laboratory (NSSL)*)

which the density depends only on the pressure. *See* FLUID-FLOW PRINCIPLES; KELVIN'S CIRCULATION THEOREM.

Vortex tubes that close on themselves roughly in the shape of a circle are not uncommon and are called vortex rings. They can readily be produced by a "vortex cannon," essentially a chamber with a circular opening opposite a drumhead. Striking the drumhead sends a puff of air that forms into a circular vortex ring outside the device. In a space without other sources of motion, a vortex ring travels surprisingly large distances: its motion is due to self-induction of the vortex lines and illustrates both induction and the notion that "vortex lines move with the fluid."

Generation of vorticity. Flows do not occur without vorticity in some region of the flow volume. Kelvin's theorem does not address the origin of vorticity. Vorticity is produced by viscosity acting at flow boundaries, by density variation, by rotation of the entire flow system, or by the action of nonconservative forces acting on the fluid. In many flows important in engineering, viscous production at boundaries is the most prevalent source of vorticity, and this often takes place only in relatively thin zones (boundary layers) adjacent to solid walls. *See* BOUNDARY-LAYER FLOW.

In natural flows that take place in the atmosphere and Earths bodies of water, vorticity is generated by viscous action too, but predominately by density variation and planetary rotation.

The last source of vorticity generation is due to action of a nonconservative force. The most prevalent nonconservative force is the electromagnetic Lorentz force. This acts in electrically conducting fluids, such as salt water, liquid metals, and ionized gases (plasmas), when electrical currents flow in the fluid in the presence of a magnetic field. This is a dominant source of vorticity in astrophysical flows, such as the sun. *See* MAGNETOHYDRODYNAMICS.

Intensification of vortices. When weak distributions of vorticity become concentrated in space we say that a vortex is formed. In flows without significant density variation or nonconservative forces, intensification occurs most commonly by one or both of two processes: vortex sheet roll-up and vortex tube stretching. Roll-up occurs because of the self-induction of vorticity elements, which causes the vortex sheet to wind increasingly tightly into tubes at the sheet edges. In this process, the vorticity in the sheet is pulled into the cores of the tubes (Fig. 2).

Vortex stretching is easiest to visualize when Kelvin's theorem applies. Then a vortex tube always encloses the same fluid material as well as the same vortex lines. If the fluid at one end of the segment of the vortex tube in Fig. 1 is moving away from the other, the tube segment stretches and its cross section decreases, a consequence of its fixed volume. The reduction of cross-sectional area squeezes the vortex lines in the bundle, increasing the vorticity in inverse proportion to the area. The speeds around the periphery of the tube increase in inverse proportion to the circumference of the tube. The increased speeds due to inward contraction of the tube is due to the same principle, conservation of angular momentum, used by a skater to increase the speed of a spin by drawing the arms inwards. *See* ANGULAR MOMENTUM.

The concentration of vorticity in a tornado and in a bathtub vortex is both due to vortex stretching. In both cases, weak background vorticity is concentrated as fluid is brought inwards. In a tornado (**Fig. 4**), this is produced by an updraft of air aloft which draws air at lower levels inwards. Draining of water produces the comparable effect in a bathtub vortex.

Sidney Leibovich

Bibliography. G. K. Batchelor, *An Introduction to Fluid Dynamics*, Cambridge University Press, 1967; H. Lamb, *Hydrodynamics*, 6th ed., 1932, reprint, Dover, 1993; H. J. Lugt, *Vortex Flow in Nature and Technology*, 1983, reprint, Krieger, 1995; P. G. Saffman, *Vortex Dynamics*, Cambridge University Press, 1992.

Vorticity

A vector proportional to the local angular velocity of a fluid element in a flowing fluid. The vorticity, $\vec{\omega}$, is a derived quantity in fluid mechanics, defined, for a flow field with velocity $\vec{u}$, by Eq. (1). As the curl of

$$\vec{\omega} = \nabla \times \vec{u} \tag{1}$$

the velocity vector, the vorticity is a vector with the dimensions of both a frequency and an angular velocity: $[\vec{\omega}] = [s^{-1}]$. As shown below, the component of vorticity along a particular axis is related to the rate of rotation of the fluid about the axis. For this

reason, flows for which $\vec{\omega} = 0$ are described as irrotational. *See* CALCULUS OF VECTORS; DIMENSIONAL ANALYSIS; VELOCITY.

Circulation. A property that is closely related to vorticity is the fluid circulation, Γ, defined, for any closed contour, C, in a fluid, by Eq. (2). In this definition, $\oint_C$ indicates the conventional counterclockwise contour integral around the contour C, $\vec{l}$ is a unit vector tangent to the contour, S is an arbitrary curved surface bounded by the contour C, and $\vec{n}$ is a unit vector normal to this surface. [The equality of the two integrals in Eq. (2) may be deduced by the application of Stokes' theorem.] The circulation is thus a scalar quantity equal to the integrated component of vorticity normal to the surface around which Γ is taken. Physically it represents the average fluid spin in a given area. Circulation is important because the Kutta-Joukowski law of aerodynamics states that the lift generated by a two-dimensional airfoil is $L = \varrho U \Gamma$. In this expression, ϱ is the fluid density, U is the free-stream velocity, and Γ is the bound circulation of the airfoil, defined conventionally as the negative of the definition above. As a consequence, the lift generated by an airplane wing is proportional to the circulation around it. *See* AERODYNAMICS; AIRFOIL; STOKES' THEOREM; SUBSONIC FLIGHT.

$$\Gamma = \oint_C \vec{u} \cdot d\vec{l} = \iint_S \vec{\omega} \cdot \vec{n}\, dS \qquad (2)$$

Example: solid-body rotation. Perhaps the simplest flow with vorticity is fluid in solid-body rotation. Such a flow is ultimately achieved when a cylindrical tank of water is spun about its axis (say, the z axis) at a constant rate, Ω. For this flow the vorticity is $\omega_z = 2\Omega$, which is equal to twice the angular velocity of the fluid element about the center of the turntable. This angular velocity is also precisely the angular velocity of each infinitesimal fluid element about its own axes, because the fluid is in solid-body motion.

The solid-body rotation flow is not the only one for which the vorticity is twice the angular velocity of a fluid element. For any flow, $\vec{\omega} \cdot \vec{a}$ (the component of $\vec{\omega}$ along the direction $\vec{a}$, where $\vec{a}$ is an arbitrary unit vector) is twice the mean angular velocity of any two line segments, through the point where $\vec{\omega}$ is evaluated, that form an orthogonal coordinate system with $\vec{a}$. For the solid-body rotation flow, the circulation around a circle of radius r centered on the origin is $\Gamma = \pi r^2 \times (2\Omega)$. Consistent with the circulation-vorticity relationship explained above, the circulation calculated here is indeed the area of the circle (πr^2) times the uniform vorticity (2Ω).

Example: irrotational vortex. Consider a flow with only a tangential velocity component, of the form $v_\theta = \Gamma/2\pi r$. Any closed loop taken around the origin of such a flow has the constant circulation Γ. Any closed loop that does not include the origin has $\Gamma = 0$. In other words, this velocity field represents a flow with infinite vorticity at the origin but zero vorticity elsewhere; i.e. the vortex, called an "irrotational vortex," is irrotational everywhere except at the origin. Such a flow is physically unattainable because the action of viscosity prohibits the infinite velocity gradients of this flow. The irrotational vortex is a mathematically useful concept, however, and forms a key element of potential flow theory. *See* POTENTIAL FLOW.

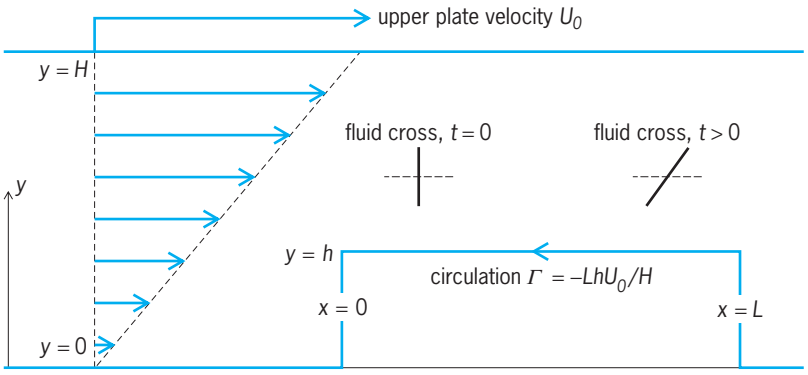


Fig. 1. Vorticity and circulation in a laminar plane Couette flow.

Example: laminar plane Couette flow. Another flow that sheds light on the concept of vorticity is the laminar Couette flow between two plates. Consider two infinite plates parallel to the x–z plane: a stationary one at $y = 0$ and a second one moving at velocity U_0 in the x direction, located at $y = H$ (**Fig. 1**). The velocity between the plates is given by $u = U_0(y/H)$, and the streamlines of the flow are thus all parallel to the x axis. The vorticity is given by $\omega_z = -U_0/H$, a constant. Surprisingly, in this parallel flow there is vorticity. The vorticity is a consequence of the fact that a cross placed in this flow would have its horizontal arm remain horizontal, whereas its vertical arm would be rotated clockwise. *See* LAMINAR FLOW.

It is instructive to consider a contour in this flow bounding the region $0 < x < L$, $0 < y < h$. The circulation about this contour is $\Gamma = -Lh(U_0/H)$. As was true for the solid-body-rotation flow, which also had constant vorticity throughout the domain, the circulation here is equal to the area of the surface (Lh) about which the circulation is measured, times the vorticity.

Vortex line and vortex tube. With a sense of what vorticity is, it is possible to contemplate what produces it and what causes it to change. It is convenient to start by defining a vortex line as a line that is everywhere tangent to the local vorticity vector (analogous to a streamline). A series of adjacent vortex lines is referred to as a vortex tube. The first Helmholtz vortex law states that at any instant in time the circulation about all loops taken around the exterior of a vortex tube is the same. Thus, vortex tubes must either form loops entirely within a fluid or terminate at some fluid boundary. *See* VORTEX.

Kelvin's theorem. Kelvin's theorem considers how the circulation Γ around a material loop in a fluid (a loop that moves with the fluid) varies in

time. Starting with the Navier-Stokes equations, Lord Kelvin showed that if (1) the flow is inviscid along the loop, (2) the fluid is subject only to potential body forces, and (3) the fluid pressure is a function of density alone, then the rate of change of Γ is 0. In other words, the circulation around a material loop is time-independent. Kelvin's theorem may also be stated slightly differently: subject to the above three constraints, vortex lines are material lines, convected with the local fluid velocity. *See* KELVIN'S CIRCULATION THEOREM; NAVIER-STOKES EQUATION.

Kelvin's theorem provides much insight into vorticity. First, consider an incompressible swirling flow in a converging section of a tube. Neglecting the three confounding effects mentioned above, as the flow converges the circulation about a material loop taken around the tube circumference remains constant. Since the flow area is diminished, this constancy of circulation implies that the mean vorticity is increased—the fluid spins more rapidly. The physical explanation for this enhancement of vorticity is that the fluid in the converging tube is stretched axially (to conserve fluid mass). This axial stretching increases the rotation rate of the fluid, for the same reason that a figure skater spins more rapidly with arms stretched overhead. The very strong winds associated with tornadoes are a result of the stretching of vortex lines in the atmosphere caused by a combination of updrafts and wind shear. *See* TORNADO.

Now, consider a cylindrical tank of air, spun about its axis and in solid-body rotation, that is gradually compressed by reducing the radius of the tank. Subject to the three constraints above, Kelvin's theorem shows that since the circulation is constant and the flow area is reduced, the vorticity is intensified. The vorticity is intensified here in order to preserve the angular momentum of the fluid. This intensification is important in engines, where, during the compression stroke, vorticity oriented parallel to the piston face (referred to as tumble) increases. *See* ANGULAR MOMENTUM; ENGINE.

Generation. Kelvin's theorem can tell us what happens when vorticity is already present in a flow, but it sheds no light on how vorticity is generated. To answer this question, it is useful to consider situations for which Kelvin's theorem is inapplicable: flow with viscosity, with nonpotential body forces, and for which the pressure is not solely a function of the density.

The action of viscosity has two effects on vorticity. One effect of viscosity is to cause the diffusion of vorticity in a fluid. The diffusion of vorticity is related to the transfer of angular momentum from one rotating object to another, by means of a frictional interface between the bodies. A classic example of this diffusion of vorticity is the Lamb-Oseen vortex. The Lamb-Oseen vortex is a two-dimensional time-dependent flow. For simplicity, one may imagine that the flow is in the x-y plane. The flow is initially an irrotational vortex of circulation Γ_o, in which the fluid travels in concentric circles at speeds inversely proportional to the distance from the center of flow. The vorticity is then zero except at the exact center of the vortex, where it is infinite. At times $t > 0$ viscosity is allowed to act. Starting with the Navier-Stokes equations; assuming that the fluid is incompressible, has constant density, and is subject only to potential body forces; and taking into account the two-dimensional axisymmetric character of the flow, it can be shown that the vorticity at times $t > 0$ and distance r from the center of flow is given by Eq. (3), where ν is the kinematic viscosity. Subjected

$$\omega_z = \frac{\Gamma_o}{4\pi\nu t} \exp\left(\frac{-r^2}{4\nu t}\right) \tag{3}$$

to viscosity, the flow develops a solid-body-rotation region, with constant vorticity, near the vortex center. The effect of vorticity is to cause the size of this region to increase, thus diffusing the initially concentrated vorticity (**Fig. 2**). The corresponding tangential velocity profiles around the vortex of Fig. 2 are shown in **Fig. 3**. *See* VISCOSITY.

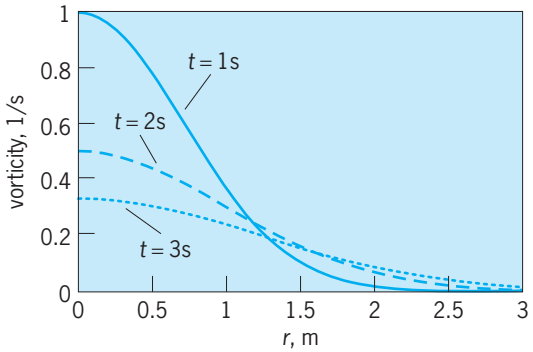


Fig. 2. Vorticity distribution in a Lamb-Oseen vortex. Here, $4\nu = 1$ and $\Gamma_o = \pi$ in SI units.

A second effect of viscosity is the generation of vorticity at a wall where there is a pressure gradient at the wall. Vorticity present in the boundary layer over a flat plate is entirely vorticity that has been generated at the leading edge of the plate, where a pressure gradient exists, which has been convected downstream.

A common example of a nonpotential body force is the Coriolis force, which is present in a rotating frame of reference. This force generates vorticity in a fluid, and is a major cause of the large-scale circulation in the atmosphere and oceans. *See* CORIOLIS ACCELERATION.

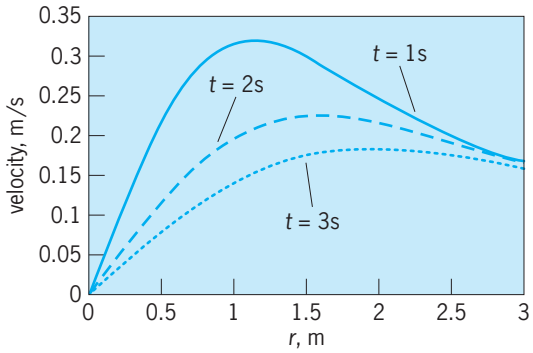


Fig. 3. Velocity distribution in a Lamb-Oseen vortex.

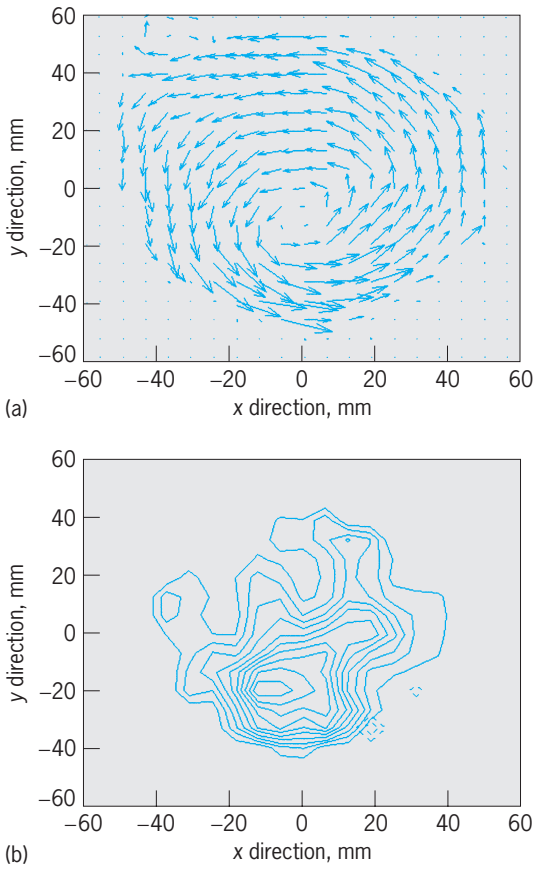


Fig. 4. Velocities and vorticity for the flow in a cardiac assist device. (*a*) Vector plot of velocities, generated by analyzing a particle image velocimetry (PIV) image taken parallel to one surface of the device. (*b*) Contour plot of the vorticity. Two small regions of negative vorticity are shown as broken contours. 1 mm = 0.04 in. (*Doug Hart and Hayden Huang, Massachusetts Institute of Technology*)

There are many flows for which the pressure may not be solely a function of the density (so-called baroclinic flows), such as the flow of gas with heat addition and the flow of water with salinity variations. Pressure gradients in such flows generate vorticity. This source of vorticity is called baroclinic torque, and is important in atmospheric flow, buoyancy-driven flow, and oceanographic flow. The physical mechanism underlying baroclinic torque is similar to that which produces a righting moment on a sailboat that is heeled over. *See* BAROCLINIC FIELD; DYNAMIC INSTABILITY; DYNAMIC METEOROLOGY.

Measurement. It is virtually impossible to measure the vorticity in a flow directly. Rather, one normally measures fluid velocities at many locations in space and then takes spatial derivatives of the velocities (since the vorticity is the curl of the velocity) to find the vorticity. Owing to the fact that derivatives of the velocity must be determined, accurate velocity information is required at a number of locations in space. Such information is commonly obtained in steady flows by means of laser Doppler velocimetry (LDV), hot-wire anemometry (HWA), or particle image velocimetry (PIV), but in unsteady flows only PIV may typically be used. **Figure 4***a* shows an example of such an unsteady flow in a cardiac assist device. Typical velocities in this device are about 50 mm/s (2 in./s). A strong vortex is present, centered slightly below and to the left of the center. Illustration *b* shows the vorticity distribution in this flow, which was computed using the data of illus. *a*. The region of highest vorticity corresponds with the center of the vortex.

Computational fluid dynamics. The governing equations of fluid mechanics may be cast in the form of differential equations for the vorticity. These are, in some (fairly rare) situations, more readily solved than the original governing equations. Since vortices form an important aspect of many fluid flows, methods that track the motion of vortices (such as large eddy simulation) have been highly successful in flow modeling. Since circulation is central to aerodynamic lift, numerical methods that treat lifting surfaces as a number of vortices (such as the vortex lattice method) can be used to model the aerodynamics of planes. *See* COMPUTATIONAL FLUID DYNAMICS; FLUID-FLOW PRINCIPLES.

Sheldon I. Green

Bibliography. S. I. Green (ed.), *Fluid Vortices*, Kluwer, Boston, 1995; A. J. Majda and A. L. Bertozzi, *Vorticity and Incompressible Flow*, Cambridge University Press, New York, 2002; P. S. Marcus, Jupiter great red spot and other vortices, *Annu. Rev. Astron. Astrophys.*, 31:523–573, 1993; P. H. Renard et al., Dynamics of flame/vortex interactions, *Prog. Energy Combust. Sci.*, 26(3):225–282, 2000; V. J. Rossow, Lift-generated vortex wakes of subsonic transport aircraft, *Prog. Aerosp. Sci.*, 35(6):507–660, 1999; T. Sarpkaya, A critical review of the intrinsic nature of vortex-induced vibrations, *J. Fluids Struct.*, 19(4):389–447, May 2004; C. H. K. Williamson and R. Govardhan, Vortex-induced vibrations, *Annu. Rev. Fluid Mech.*, 36:413–455, 2004.

Wage incentives — Wurtzilite

Wage incentives

Plans that link employee compensation to some measure of company success. Wage incentive plans throughout the United States are varied. Some plans endeavor to link compensation to the company's performance history, others to department or unit performance, others to individual or team success. However, the majority of wage incentive plans are linked to individual achievement.

Modern wage incentive systems provide extra earnings when achievement exceeds some established bench mark. However, workers receive only a base rate, which in some installations may be reduced when achievements become less than the established bench marks. Thus, modern wage incentive plans promote the concept that workers share in both the rewards and the risks in doing business.

Experience has proven that properly designed and managed wage incentive plans that tie pay closely to performance are a way that companies can increase financial rewards and achieve gains in both productivity and quality. Plans must be developed so that workers understand that their individual efforts are directly related to the established benchmark goals. Specific incentive plans attached to individuals or small groups such as teams rather than divisions or whole plants usually are more effective.

Established benchmarks often include more than a given number of pieces produced. Such factors as product quality, plant safety, and customer satisfaction are involved in establishing the wage incentive plan. A plant that applies wage incentives throughout the organization will usually have a different plan for services than it has for direct labor, where output, quality, and safety readily are measurable.

The following criteria help assure the success of the wage incentive installation.

1. *Benchmark goals.* These should be established by careful measurement and should be attainable. Workers should understand how the benchmarks were established and realize that the workers can readily meet the targets when the members of their group have been properly trained and are giving normal effort.

2. *Meaningful goals*. Quantity goals are easily understood. However, goals based on quality must be based upon readily understood targets such as "no more than 1% defective parts per 1000" or "no more than one day of lost time due to accidents per month."

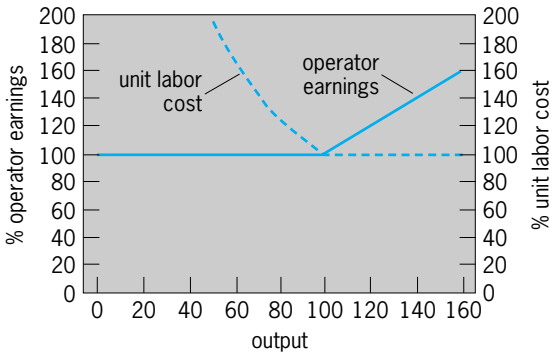
3. *Employee participation*. As the wage incentive plans are being developed, it is important to give employees an opportunity to make suggestions.

4. *Flexibility*. It is important to adjust plans as necessary in order to take into account changing methods, facilities, products, and customers.

The principal wage-incentive plans are piecework, standard hour, gain sharing, and profit sharing.

Piecework. Implied in the term piecework is the concept that all standards are expressed in money and that workers are rewarded in direct proportion to output. Under straight piecework, the day rate is not guaranteed. Since in the United States federal law requires a minimum guaranteed hourly rate for full-time employees, straight piecework is not used on a large scale. The reasons for the popularity of piecework are that it is easily understood by the worker, easily applied, and easily kept up to date as methods or product mix change. It is widely used in the fruit and farm harvest industry. When used in conjunction with the minimum wage laws, the base rate is guaranteed, and this is usually established by job evaluation. Output of the worker to the point of the guaranteed base rate usually is made a requirement of employment, and output above the guaranteed base rate is compensated in direct proportion to work produced.

Standard hour plan. The standard hour plan with a guaranteed base rate, established by job evaluation, is a form of incentive wage payment that is used frequently. The difference between the standard



Relationship of labor cost and operator earnings to operator output.

hour plan and piecework is that under the standard hour plan, standards are expressed in time rather than money (see **illus.**). Operator earnings are based upon standard hours earned multiplied by base rate per hour. If the worker earns fewer standard hours than clock hours, the wage paid is calculated as clock hours times base rate per hour.

Certain requisites for a sound standard hour plan have been established. It has been proven that a sound standard hour wage-incentive plan will increase the rate of production with no detrimental effect of quality, lower overall unit costs, reduce supervision costs, and promote increased earnings of employees. Before installation of a wage-incentive plan based upon output is made, there should be incorporated a policy of methods standardization so that sound work measurement can be accomplished. Also, established base rates should be equitable and should provide for a sufficient spread between job classes to recognize the positions that demand more skill, effort, and responsibility.

Fair standards of performance must be developed before a valid standard hour installation can take place. With both piecework and standard hour plans, standards should be based on some form of work measurement such as time study, fundamental motion data, standard data, formula, or work sampling. The plan should give workers the opportunity to earn approximately 20–35% above their base rate if they are normally skilled, are properly trained, and execute high effort continually. The terms of the plan should be understood easily by all who are compensated by it. The plan should guarantee the basic hourly rate set by job evaluation; this basic rate should be a good living wage comparable to the prevailing wage rate of the area for each job in question. There should be a range of pay rates for each job, and these should be related to total performance—a concept that includes quality, reliability, safety, and attendance as well as output. *See* WORK MEASUREMENT.

Sound standard hour wage-incentive systems have proven to be one of the most effective motivators to increase productivity. Productivity, which is the measure of output based upon following a prescribed method and utilizing average skill and effort, is assigned a value of 100%. This value frequently is increased to 135%, and sometimes a higher percentage, under a well designed and administered standard hour incentive installation.

In those organizations where a detailed methods study has not been made and with no work measurement or wage incentives, productivity typically is about 50%. With methods analysis followed by work measurement and good supervision, the level of productivity usually rises to 90–100%. A standard hour wage-incentive installation can improve this to 125–135%. *See* PRODUCTIVITY.

Gain-sharing plans. These plans are also referred to as productivity sharing plans. Gain-sharing plans are characterized by sharing the benefits of improved productivity, cost reduction, quality improvement, customer satisfaction, and accident reduction. These plans are often used in addition to some other plan such as straight day work or a standard hour plan. Gain-sharing plans have become popular, since there are several factors other than productivity that can enhance the success of an industry or business.

Gain-sharing incentive plans can provide motivation for savings in material (both direct uses and factory supplies), reduction in rejects and scrap, improvement in product quality, accident prevention, and improved customer satisfaction. The plans tie the incentive to bench marks. Performance that achieves results better than established benchmarks is converted into monetary values that are shared by workers and the company. Benchmark values are usually based upon average historical performance for a period of time, usually one year. Often the bench mark is a production value, which is the monetary difference between sales and purchases, and consequently represents the value that all employees have contributed. Previous performance, for the past year, is considered standard. Value added can be improved by (1) savings in raw materials, purchased parts, supplies, fuel, and power; (2) reduction in rejects, rework, and scrap; (3) improvement in product overall quality; (4) reduction in worker absenteeism; (5) reduction in customer allowances, including less warranty service; (6) increase in the volume of output; (7) reduction in lost time resulting from accidents; and (8) greater productivity without increases in time inputs by hourly or salaried employees.

Under gain-sharing plans, management usually computes incentives on a monthly basis, and this is identified as a separate check from the regular paycheck (typically issued weekly). Customarily, only two-thirds of the incentive earned in a given pay period is distributed. The remaining third is placed in a reserve fund to be used any month that performance falls below standard.

Profit sharing. Profit sharing may be thought of as a form of gain sharing, since it refers to any procedure where an employer pays to all employees, in addition to the standard rates of regular pay, special current or deferred sums based not only upon individual or group performance but on the prosperity of the business as a whole. Most profit-sharing systems are either cash or deferred plans. Cash plans give

periodic distributions of money from the profits of the business, while deferred plans feature the periodic investment of portions of the profits for employees. A successful profit-sharing program depends on the profits of the company, which frequently are not under the control of the labor force. In periods of low profits or of losses, the plan may actually weaken rather than strengthen employee morale. *See* WORK STANDARDIZATION. Benjamin W. Niebel

Bibliography. C. H. Fay and R. W. Beatty, *The Compensation Source Book*, 1988; H. Gleckman, S. Atchison, and T. Smart, Bonus pay: Buzzword or bonanza, *Bus. Week*, pp. 62–64, November 14, 1994; J. Hill, Employees blunder into rewards, *U.S.A. Today*, May 22, 1990; B. W. Niebel, *Motion and Time Study*, 9th ed., 1993; G. Salvendy (ed.), *Handbook of Industrial Engineering,* 3d ed., 2001.

Flow visualization of aircraft wing-tip vortices. (*NASA*)

Wake flow

The flow downstream of a body immersed in a stream, or the flow behind a body propagating through a fluid otherwise at rest. Wakes are narrow elongated regions aligned with the flow direction and filled with large and small eddies. The wake eddies of a bridge pier immersed in a river stream, or of a ship propelled through the water, are often visible on the surface. On windy days, similar wakes form downwind of towers, smoke stacks, or other structures, but such eddies in the air are not generally visible unless some smoke or dust is entrained in them. Wakes are sustained for very large distances downstream of a body. Ship wakes retain their turbulent character for miles behind a vessel and can be detected by special satellites hours after their generation. Similarly, condensation in the wake of aircraft sometimes looks like narrow braided clouds, traversing the sky.

Turbulence in the wake of bluff bodies consists of all sizes of eddies, which interact with each other in their unruly motion. Yet, out of this chaos emerges some organization, whereby large groups of eddies form a well-ordered sequence of vortices. These vortices are rolling and moving downstream much like the waves on the surface of the ocean, and for this reason they are often referred to as rollers. The sense of rotation of these vortices alternates, and their spacing is quite regular. As a result, they can drive a structure that they encounter, or they can exert on the body that created them a force alternating in sign with the same frequency as that of the formation of the vortices. Such rollers can be experienced when following in the wake of a large truck. *See* KÁRMÁN VORTEX STREET; TURBULENT FLOW.

The flow over bodies that generate steady and significant lifing forces, that is, nonbluff bodies, forms wakes dominated by vortices that are aligned with the direction of the oncoming stream. Such vortices are often called streamers. Typical streamers are the tip vortices that are shed from the tips of airplane wings (see **illustration**). Tip vortices pose grave danger to following aircraft, especially in cases of landing. These vortices continue spinning over the landing area minutes after an aircraft has landed and taxied away, and could induce violent motions in another aircraft with catastrophic results. This phenomenon extends the time between landings and is the major cause of airport congestion. *See* WING.

The same type of streamwise vortices is encountered over other types of engineering structures with different effects. For example, streamers are shed off the corners of buildings during a violent storm, creating very large suction forces that damage the roofing material or lift the entire roof. *See* FLUID FLOW; FLUID-FLOW PRINCIPLES; VORTEX. Demetri P. Telionis

Bibliography. R. L. Panton, *Incompressible Flow*, Wiley Interscience, 3d ed., 2005; F. M. White, *Fluid Mechanics*, McGraw-Hill, 5th ed., 2003.

Wall construction

Methods for constructing walls for buildings. Walls are constructed in different forms and of various materials to serve several functions. Exterior walls protect the building interior from external environmental effects such as heat and cold, sunlight, ultraviolet radiation, rain and snow, and sound, while containing desirable interior environmental conditions. Walls are also designed to provide resistance to passage of fire for some defined period of time, such as a one-hour wall. Walls often contain doors and windows, which provide for controlled passage of environmental factors and people through the wall line.

Structure and design. Walls are designed to be strong enough to safely resist the horizontal and vertical forces imposed upon them, as defined by building codes. Such loads include wind forces, self-weight, possibly the weights of walls and floors from above, the effects of expansion and contraction as generated by temperature and humidity variations as well as by certain impacts, and the wear and tear of interior occupancy. Some walls are also designed to provide

diaphragm strength to help keep the entire building erect. Diaphragm strength refers to the ability of a wall to resist forces applied in the plane of the wall. Thus, a design to resist wind or earthquake forces may utilize walls with diaphragm strength, oriented parallel to such forces. If a wall has windows or doors, they require frame and lintel structure within the wall itself. *See* LOADS, DYNAMIC; LOADS, TRANSVERSE.

Older walls typically had to help provide support for the floors and walls above, and therefore they were classified as bearing walls. Modern engineering has led to lighter buildings that can be supported efficiently by structural frames of steel, concrete, or wood, instead of by the walls. This has led to the development of curtain walls, which are nonbearing and provide primarily for enclosure, not for structural support of the building.

Types of modern walls. Modern building walls may be designed to serve as either bearing walls or curtain walls (**Fig. 1**) or as a combination of both (**Fig. 2**) in response to the design requirements of the building as a whole. Both types may appear similar when complete, but their sequence of construction is usually different.

Bearing walls. Bearing-wall construction may be masonry, cast-in-place or precast reinforced concrete, studs and sheathing, and composite types. The design loads in bearing walls are the vertical loading from above, plus horizontal loads, both

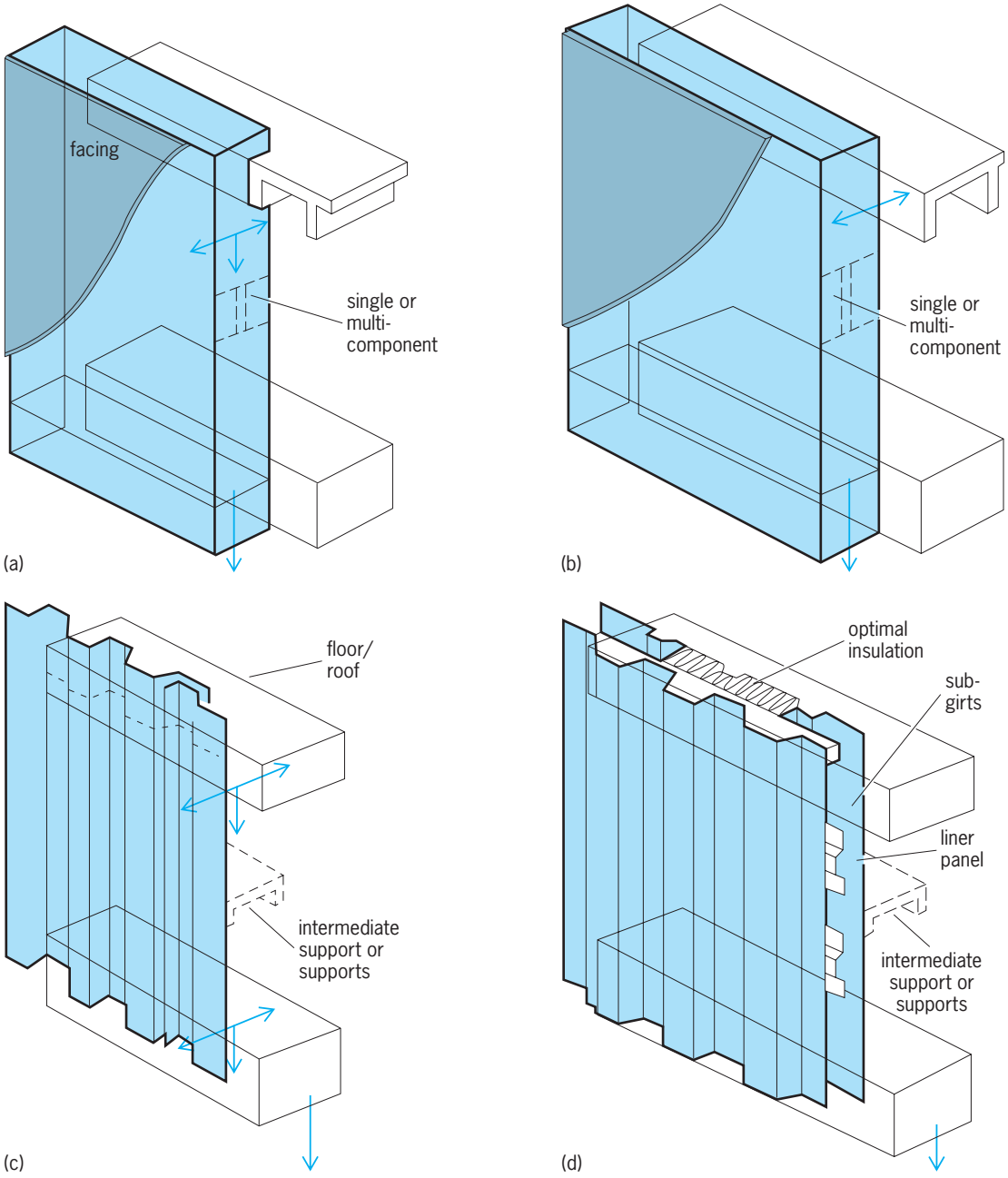


Fig. 1. Basic bearing- and curtain-wall construction. Arrows refer to force vectors, such as vertical dead loads, or horizontal wind loads. (*a*) Bearing wall. (*b*) Nonbearing (but self-supporting) wall. (*c*) Formed metal curtain wall, single thickness. (*d*) Formed metal sandwich curtain wall. (*After Sweets Catalog File: General Building and Renovation, 1993*)

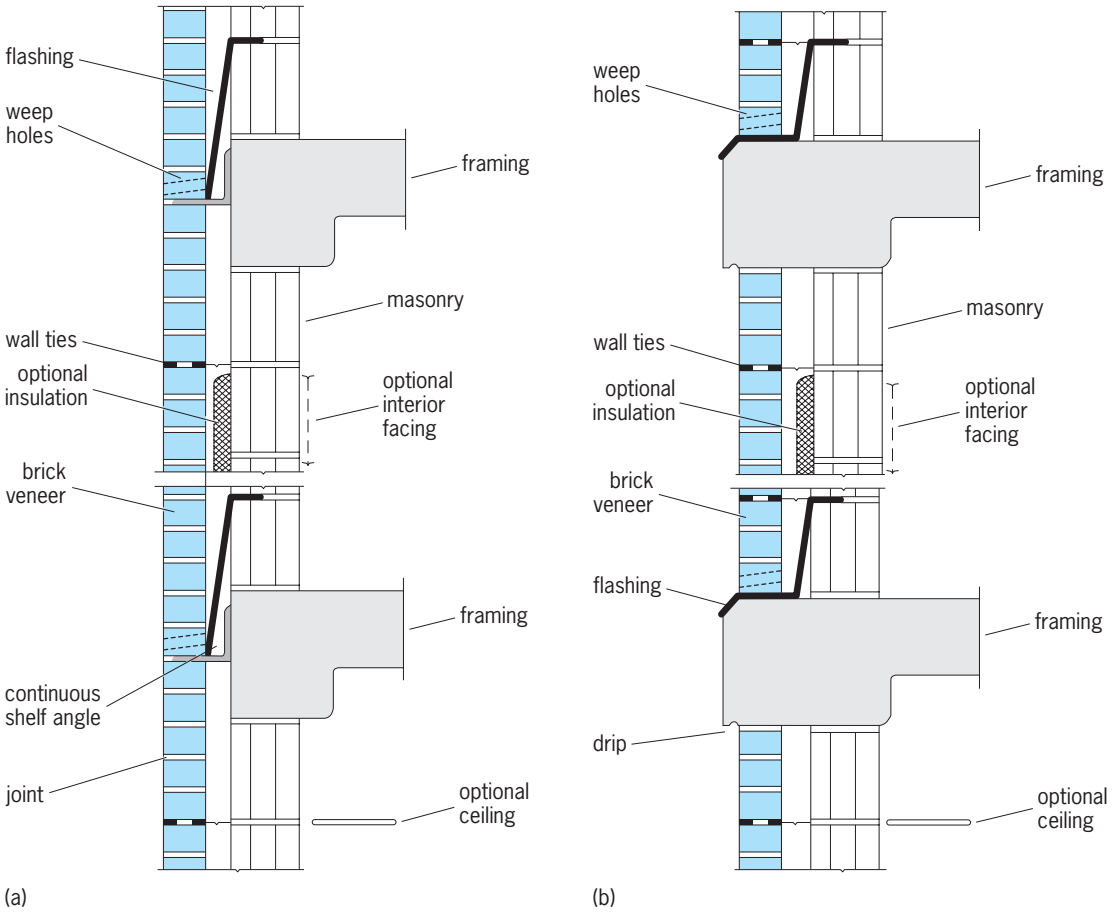


Fig. 2. Details of some masonry wall construction. (*a*) Continuous veneer. (*b*) Interrupted veneer. (*After Sweets Catalog File: General Building and Renovation, 1993*)

perpendicular and parallel to the wall plane. Bearing walls must be erected before supported building components above can be erected.

Curtain walls. Curtain-wall construction takes several forms, including lighter versions of those used for bearing walls. These walls can also comprise assemblies of corrugated metal sheets, glass panels, or ceramic-coated metal panels, each laterally supported by light subframing members. The curtain wall can be erected after the building frame is completed, since it receives vertical support by spandrel beams, or relieving angles, at the wall line (Fig. 2). The primary design loading for a curtain wall is lateral loading, perpendicular to the wall, due to wind and interior occupancy.

Masonry walls. These are a traditional, common, and durable form of wall construction used in both bearing and curtain walls. They are designed in accordance with building codes and are constructed by individual placement of bricks, blocks of stone, cinder concrete, cut stone, or combinations of these. The units are bonded together by mortar, which is composed of cement, sand, water, and sometimes lime; the mortar may be chemically enhanced by latex or epoxy additives to provide greater tensile strength. Sometimes hollow-block masonry walls are reinforced by vertical rebars, grouted into the block openings. *See* MORTAR.

Masonry walls are usually built up of at least two separate thickness, called wythes. An example is a wall composed of 4-in. (10-cm) brick and 8-in. (20-cm) concrete block. The wythes may be bonded together to form a solid wall, or they may have a space between them to form a cavity wall. The two wythes are usually structurally connected so as to act essentially as a single structural unit.

In the solid wall, the wythes are bonded to each other by mortar. Such solid walls commonly offer little insulation value, and may allow some moisture penetration from wind-driven rains unless special veneers or coatings are provided on the exterior or interior surfaces.

In the cavity wall, the wythes are typically connected to each other across the cavity by spaced steel wall ties or by welded steel trussed or parallel rod reinforcement, embedded in opposite mortar joints in each wythe. Special ties may be used to permit small movements between the two wythes.

The cavity is often used for inclusion of insulation such as fiberglass, styrofoam, or urethane foam; it also allows for drainage of any condensed or leaked moisture. Such moisture falls downward by gravity and is intercepted by waterproof metal, plastic, or asphalted fabric flashing, which spans across and downward through the cavity. This flashing causes such water to leave the wall through weep holes,

which are small holes placed in the mortar joints and spaced several feet apart (Fig. 2).

Some masonry walls, as well as other types of walls, may have a veneer of special materials such as thin brick; ceramic tile; or cementitious, elastomeric, or other water-resistant coatings; these are affixed to the exterior or interior faces for esthetic or protective reasons. *See* CONCRETE; MASONRY.

Reinforced concrete walls. These are used for both strength and esthetic purposes. Such walls may be cast in place or precast, and they may be bearing or curtain walls. Some precast concrete walls are constructed of tee-shaped or rectangular prestressed concrete beams, which are more commonly used for floor or roof deck construction. They are placed vertically, side by side, and caulked at adjacent edges. Sometimes the casting encloses insulation. In other solid types, insulation is attached on the inside face of the casting and is then protected by an inside facing material. *See* CONCRETE BEAM; REINFORCED CONCRETE.

Stud and sheathing walls. These walls are a light type of wall construction, commonly used in residential or other light construction where they usually serve as light bearing walls. They usually consist of wood sheathing nailed to wood or steel studs, usually with the dimensions 2 × 4 in. (5 × 10 cm) or 2 × 6 in. (5 × 15 cm), and spaced at 16 in. (40 cm) or 24 in. (60 cm) on center—all common building module dimensions.

The interior sides of the studs are usually covered with an attached facing material. This is often sheetrock, which is a sandwich of gypsum between cardboard facings. Insulation is usually enclosed within the stud wall cavities. A vapor barrier may also be provided; this is a coating or a membrane that minimizes humidity penetration and possible condensation within the wall.

Composite walls are essentially a more substantial form of stud walls. They are constructed of cementitious materials, such as weatherproof sheetrock or precast concrete as an exterior sheathing, and sheetrock as an interior surface finish. *See* GYPSUM; PRECAST CONCRETE.

Prefabricated walls. These are commonly used for curtain-wall construction and are frequently known as prefab walls. Prefabricated walls are usually made of corrugated steel or aluminum sheets, although they are sometimes constructed of fiber-reinforced plastic sheets, fastened to light horizontal beams (girts) spaced several feet apart (Fig. 1). The corrugations usually span vertically to deliver the lateral wind loading to the horizontal girts, which then carry such loading to the primary vertical framing structure. Prefab walls are often made of sandwich construction: outside corrugated sheets, an inside liner of flat or corrugated sheet, and an enclosed insulation are fastened together by screws to form a thin, effective sandwich wall. Sometimes the inner and outer sheets and the enclosed insulation are factory-fabricated by adhesion to form a very strong and stiff composite sandwich wall plank. These usually have tongue-and-groove vertical edges to permit sealed joints when the units are erected at the building site by being fastened to framing girts.

Panel walls. Glass, metal, or ceramic-coated metal panel walls are a common type of curtain wall used in high-rise construction. They are typically assembled as a sandwich by using glass, formed metal, or ceramic-coated metal sheets on the outside, and some form of liner, including possibly masonry, on the inside; insulation is enclosed. Sometimes such walls include masonry and insulation, about 3 ft (0.9 m) high, to form a fire-resistant wainscot, with the upper section of only glass, in one or more plys. Some modern building walls are made predominantly of glass sheets alone, fastened or adhered at the edges by elastomer adhesive/sealants.

Other types. Tilt-up walls are sometimes used for construction efficiency. Here, a wall of any of the various types is fabricated in a horizontal position at ground level, and it is then tilted up and connected at its edges to adjacent tilt-up wall sections. Interior partitions are a lighter form of wall used to separate interior areas in buildings. They are usually nonbearing, constructed as thinner versions of some of the standard wall types; and they are often designed for some resistance to fire and sound. Retaining walls are used as exterior walls of basements to resist outside soil pressure. They are usually of reinforced concrete; however, where the basement depth or exterior soil height is low, the wall may be constructed as a masonry wall. *See* BUILDINGS; RETAINING WALL; STRUCTURAL MATERIALS.

Milton Alpern

Bibliography. American Society of Civil Engineers, *Building Code Requirements for Masonry Structures*, 1990; American Society of Civil Engineers, *Minimum Design Loads for Buildings and Other Structures*, 1990; D. K. Ballast, *Architect's Handbook of Construction Detailing*, 1990; C. Beall, *Masonry Design and Detailing for Architects, Engineers and Contractors*, 4th ed., 1997; H. J. Cowan, *Design of Reinforced Concrete Structures*, 2d ed., 1988; K. F. Faherty and T. G. Williamson (eds.), *Wood Engineering and Construction Handbook*, 3d ed., 1998.

Walnut

This name is applied to about a dozen species of large deciduous trees widely distributed over temperate North and South America, southeastern Europe, and central and eastern Asia. The genus (*Juglans*) is characterized by pinnately compound aromatic leaves and chambered or laminate pith (**Fig. 1**). The staminate (male) flowers are borne in unbranched catkins on the previous season's growth, and the pistillate (female) flowers are terminal on the current season's shoots. Pollination is by wind. The shells of the nuts of most species are deeply furrowed or sculptured.

The plants fall into two natural groups, one characterized by a prominent band of hairs on the upper edge of the leaf scar, pointed buds, and usually elongate nuts. The leaf scars of the other group lack the cushion of hairs, and the nuts are spherical or nearly

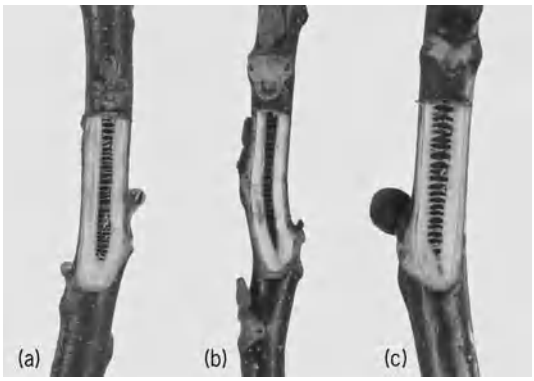


Fig. 1. Pith of (*a*) black walnut (*Juglans nigra*), (*b*) butternut (*J. cinera*), and (*c*) English walnut (*J. regia*).

Fig. 2. Branches showing buds and leaf scars of (*a*) black walnut (*Juglans nigra*), (*b*) butternut (*J. cinera*), and (*c*) English walnut (*J. regia*).

so. Of the six North American species, five are in the latter group and one, the butternut or white walnut (*J. cinera*), in the former (Fig. 1 and **Fig. 2**).

Two species, the black walnut (*J. nigra*) and the Persian or English walnut (*J. regia*), are of primary importance for their timber and nuts. The butternut finds local use in the northeastern United States. The other species are sparingly used as shade trees, as grafting stocks, and as sources of nuts.

The black walnut and other plants in the walnut family produce a toxic substance called juglone (a napthaquinone), which according to present evidence suppresses the growth of many plants when their roots come in contact with walnut roots. Tomatoes, alfalfa, rhododendron, apple, and pine are particularly affected. However, some plants, notably some of the grasses and the black raspberry, appear to be stimulated by the walnut roots. Some recent research relates laminitis in horses to their being bedded down with fresh black walnut wood chips. The chemistry of the injury is not well understood.

Black walnut. The black walnut is a large tree; it can reach a height of over 100 feet (30 m) and a diameter of 3–4 ft (0.9–1.2 m). It is native to the hardwood forests of the Central Mississippi Valley and the Appalachian region of North America.

Walnuts thrive on rich, well-drained, alluvial soils well supplied with moisture and are most abundant in the river valleys and along streams. They have, however, been planted beyond their natural range and succeed on good agricultural soils over a wide area.

The pinnately compound leaves have 15–20 leaflets, and the stamens and pistils are borne separately (monoecious) as in other members of the genus. The drupelike spherical fruit $1^1/_2$–$2^1/_2$ in. (38–64 mm) long is borne singly or in clusters of two or three. It consists of an outer indehiscent fleshy husk about $^1/_2$ in. (13 mm) thick enclosing the hard, rough-shelled nut $1^1/_4$–$1^1/_2$ in. (32–38 mm) in length. The kernels formed from the cotyledons are enclosed in membranous seed coats which turn brown if nuts remain in husks (**Fig. 3**). The dark-brown, strong, durable wood is rated as North America's best for gunstocks, fine furniture, and veneer. The annual demand for high-quality logs exceeds the supply. This is in part due to the export of walnut logs.

Nut production is mostly confined to the millions of trees in the forests and farmsteads. Formerly the nuts were gathered in large quantities, hulled, dried, and cracked for home use, or the operation was a cottage industry. With the advent of cracking machines about 1935, shelling and processing plants developed in the Central Mississippi Valley. Also, there is the beginning of a black walnut orchard industry which aims at production of both timber and nuts. Under orchard conditions, the culture and harvesting of the black walnut crop will undoubtedly be mechanized as with the Persian or English walnut.

The average kernel yield of seedling black walnuts in shell is 10–12%. Since the early 1900s the Northern Nut Growers Association has located and propagated many varieties (cultivars) with 20–30% kernel. Some of the more important of these are the Thomas, Burns, Emma K, Victoria, Stambaugh, Snyder, and Sparrow. There is a shortage of grafted trees at the present time. Although these cultivars are much better than the average wild seedlings, they have not been good enough to be commercially profitable or raised in sufficient quantity to establish a special market. They are, however, planted for home use.

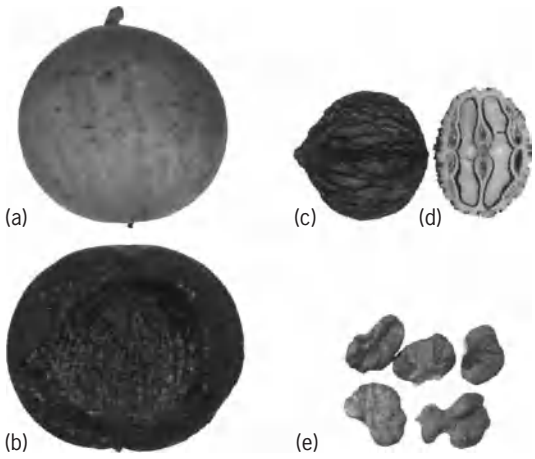


Fig. 3. Black walnut (*Juglans nigra*). (*a*) Mature fruit. (*b*) Nut with half of hull cut away. (*c*) Nut without husk. (*d*) Nut cut transversely to show shell structure. (*e*) Kernels in quarters.

Black walnut kernels were an important source of food for the American Indians and the early settlers in the United States. Now they are extensively used in ice cream, candy, and baked goods. Kernels are graded at the processing plant according to size of the pieces. In the cracking process part of the kernels are recovered as whole quarters and, with the fragments larger than $^1/_6$ in. (4 mm), are sold in plastic bags, glass jars, or metal cans. The dense hard shells of the nuts are ground and used for drilling mud in oil fields, for abrasives for polishing metal castings, and for paint fillers and other industrial uses.

English walnut. The most important of the nut-bearing trees belonging to the genus *Juglans* is the English walnut. Native to central Asia and Asia Minor, the tree was distributed widely in ancient times throughout the temperate and subtropical climatic zones of the Old World. It is also called Persian walnut, which is considered a more appropriate name. The large, round-headed, long-lived trees produce the circassian walnut lumber of commerce that is valued for gunstocks, furniture, veneer, and paneling. In France, the Balkans, and elsewhere the trees are raised for both nuts and lumber.

The leaves are pinnately compound with five to nine entire leaflets, variable in shape and size. The trees are monoecious, with the staminate flowers borne in catkins and the nutlets terminal on the current season's growth. The oval fruit, $1^1/_2$–2 in. (38–51 mm) long, in good varieties, consists of a fleshy outer hull about $^1/_4$ in. (6 mm) thick which splits irregularly at maturity, freeing the nut (**Fig. 4**). The wild-type nuts of central Asian origin are small and hard-shelled.

Through centuries of selection, large thin-shelled types have been developed, varying widely in their adaptation to different climates. Thus it is important in any commercial venture to choose varieties (cultivars) for their characteristics related to hardiness, total heat requirements, dormancy, bearing habit, and other factors. The commercial Persian walnut industry in the United States is centered in northern and central California, where the climate resembles that of the Mediterannean region where walnuts have thrived for centuries. Earlier plantings in Oregon and Washington have been reduced because of losses from winter injury. Planting has also shifted from southern California because of low yields related to failure of the trees to break dormancy, and to damage to the nuts from sunburn. The bulk of the California walnut crop is produced from orchards of older varieties such as Franquette, Hartley, Ashley, Eureka, Chico, Amigo, and Payne.

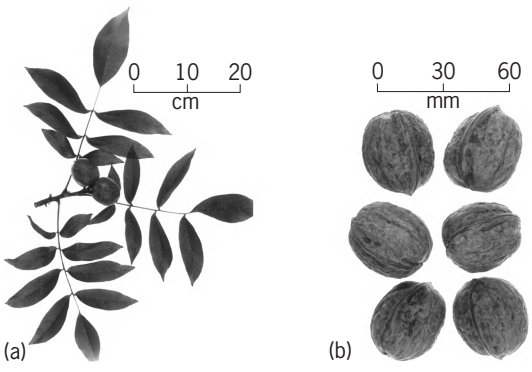


Fig. 4. English walnut (*Juglans regia*). (*a*) Twig with leaves and fruit. (*b*) Hulled nuts.

New varieties produced by breeding give greatly increased yields related to their having a high percentage of lateral buds that produce nuts. Typically in Persian walnuts pistillate flowers arise from the terminal or subterminal buds. The newer, heavy-yielding cultivars may fruit from 80% of lateral buds. Some of the cultivars with this characteristic are the Payne, Ashley, Chandler, Howard, and Sunland. Because the pistillate and staminate flowers are borne separately (monoecious) and may mature at different times (dichogamy), cultivars should be planted that will shed the wind-borne pollen at the time the pistillate flowers are receptive, either by planting cultivars whose flowers mature at the same time or by planting combinations of cultivars that shed pollen at the right time to pollinate all receptive pistils. The cultivars Amigo and Chico are used in commercial orchards particularly for their value as pollenizers. Collecting and storing pollen for artificial release at the right time is a practical procedure.

Production in the world other than in the United States and western Europe is mostly from seedling trees which receive little care. In the better orchards of California, grafted trees of high-yielding varieties are given modern highly mechanized intensive care. Propagation of the cultivars is by patch budding or whip grafting. A variety of stocks have been used. Persian walnut stock is the most compatible, but lost favor earlier because of susceptibility to root rot. California black walnut (*J. hindsii*) and hybrid (Paradox) stocks have been used, but show graft union failure from blackline disease as the trees age. The use of *J. regia* stocks is now increasing.

For good production the planting site should have good air drainage to prevent frost damage and a deep, well-drained, neutral or slightly alkaline soil. Frequent fog or rain may make disease control difficult.

Planting distance in the older orchards has been 50 ft (15 m) for mature trees. In modern orchards using early bearing cultivars which give heavy crops from lateral buds, recommended distance is 25–30 ft (7.5–9 m) with the option of removing half the trees when they begin to crowd, which will be in about 12–15 years.

Pruning is done to develop a strong framework which will support the crop. The young tree is trained with a central leader until four or five well-spaced permanent branches are developed. Then the center is opened up to form a modified leader pattern. Cultivars that bear nuts only on terminal buds should not be cut back to restrain growth. Those bearing nuts from lateral buds may be cut back without seriously affecting yield.

Good cultural practice requires weed control, cover crops, fertilization, and irrigation. An ample water supply is essential to good production. This

must be related to the highly mechanized harvesting procedures.

The more important insect pests are various husk flies, weevils, curculio, and codling moth. Troublesome diseases are walnut anthracnose, walnut blight, and various root rots. Some types of witches'-broom are related to zinc deficiency, but recent research indicates that a microplasma may be involved for which there is now no control. Appropriate pest and disease controls are essential to producing a commercial crop.

Yields of walnuts vary widely in different orchards depending on differences in site, soil, and many other factors. Fluctuations occur in overall production from year to year depending on seasonal conditions particularly related to frost, rainfall, and alternate bearing cycle. Yields of 4 tons in shell per acre (9 metric tons/hectare) have been reported, but 2–3 tons (4.5–6.75 metric tons) is more likely, and the 5-year average of all orchards is about 1 ton/acre (2.25 metric tons/hectare).

Harvesting and processing in most of the producing countries is done with local hand labor. In California, however, operations are highly mechanized. Before harvest, the ground under the trees is cleaned and leveled. Machines shake the nuts from the trees and sweep them into windrows, from which they are picked up mechanically and taken in bulk to the processing plant, where they are hulled, washed, and dried. About 25% of the crop to be marketed in the shells is sized and bleached. The others are machine-shelled, and the kernels are graded for color by electronic machines, sized, and packaged for the retail trade. The Diamond Walnut Growers, Inc., handles over 50% of the crop and determines quality and grade. The Walnut Control Board decides the amounts available for export and the proportion to be shelled.

California and Oregon produce all commercial quantities of walnuts in the United States. Other producing countries are France, Italy, and Turkey. Russia and China also produce commercial quantities of walnuts.

More than half of the English walnuts are shelled, and the kernels are used in a great variety of ways in baking and confectionery, and are also salted. The others are sold in shell to the consumer. In foreign countries many of the nuts are used locally as food.

Carpathian walnut. There is considerable interest among producers in the central and eastern United States in growing the Carpathian walnut. This strain of Persian walnuts was introduced from the Carpathian Mountains in the 1920s by Paul Crath. Prior to that time the attempts to grow Persian walnuts in the eastern United States were with varieties mostly of Mediterranean origin which failed because the trees were killed at temperatures of about 15° (−9°C). The Carpathian introductions, when dormant, withstood temperatures of −35°F (−37°C) and attracted much attention. Seeds and seedling trees were distributed widely through eastern North American from Ontario, Canada, southward into the northern parts of the Southern states. From these trees and others of outstanding hardiness, selections have been made, named, and propagated. Some of the more promising cultivars are Hansen, Colby, Broadview, Metcalfe, McKinster, and Somers. There is a concerted effort by the Northern Nut Growers Association to secure hardy varieties adapted to the central and northeastern United States from all sources. The term Carpathian walnut is now applied to these regardless of origin. Most of the trees offered by nurseries are seedlings, many of which bear nuts of poor quality, have a late bearing age, and fail to produce crops. Causes of failure may be late spring frost, lack of pollination, and pest and disease damage. There is no commercial Persian walnut industry in the eastern United States, but many trees have been planted by amateurs in a search for better varieties. There is a limited market for seed nuts; otherwise production is absorbed locally.

Butternut. The butternut, or white walnut, is the hardiest of the American species of *Juglans*, its range extending from Maine and New Brunswick, westward to Ontario, south to Arkansas and the mountains of Georgia. The trees are adapted to upland soils and in the northeastern United States are one of the few nut-bearing species. Distinguishing characters are the pointed buds and the cushion of hairs above the leaf scar (Figs. 1 and 2). The nuts, which are borne in clusters of two or three, are ellipsoid and pointed and are enclosed in a tight indehiscent husk which is covered with sticky hairs. The shell is deeply sculptured with jagged ridges.

The butternut, as compared with the black walnut, has a shallow root system and a more spreading crown. Trees are often short-lived because of defoliation by fungus disease. The wood is lighter-colored and less dense than the black walnut, and is used for interior finish and furniture. The husks of the nuts are formerly used in dying cloth greenish yellow.

Butternuts are used locally for their highly flavored oily kernels. Nuts from most of the wild trees are difficult to crack. Some progress has been made in selecting clones with good cracking quality.

Laurence H. MacDaniels

Diseases. Black walnut, English walnut, and butternut are susceptible to a number of leaf, stem, and root diseases. The seriousness of specific diseases varies greatly with tree species. Walnut anthracnose, caused by the fungus *Gnomonia leptostyla*, with the imperfect stage *Marssoninna juglandis*, is the most serious foliar disease of black walnut. The anthracnose fungus also attacks butternut, English walnut, Hinds walnut, and California walnut. Symptoms appear on the leaflets as small, circular to irregularly shaped, brown to black areas, often surrounded with a yellowish halo (**Fig. 5**). Two or more spots may later merge to form a large blighted area. Affected leaflets may turn brown and drop early. Disease losses can be reduced by the application of nitrogen fertilizers and protective sprays of the fungicide benomyl.

Two other foliar pathogens of black walnut cause diseases of lesser importance. The fungus *Microstroma juglandis* causes a white mold or a yellow leaf blotch. The fungus *Cristulariella pyramidalis*

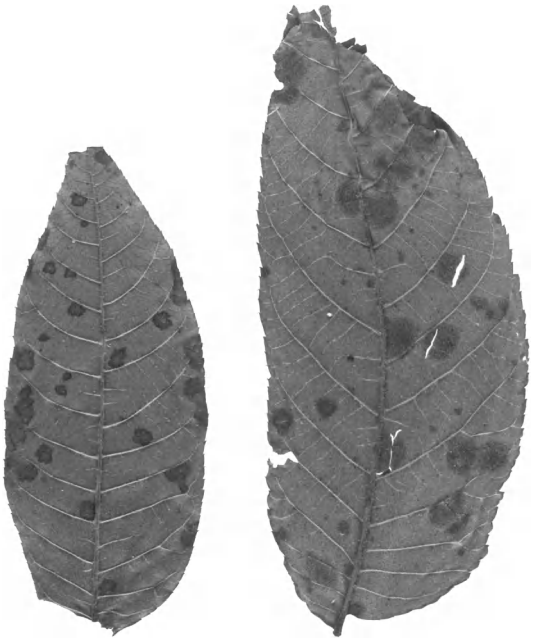


Fig. 5. Walnut leaves showing anthracnose symptoms.

has caused severe premature defoliation of trees in Illinois and Ohio.

Bacterial blight, caused by *Xanthomonas juglandis*, is especially severe on English walnut. Buds, leaves, and shoots are attacked, but the disease is most serious and destructive when quality and yield of nuts are reduced. Small reddish-brown spots occur on the leaves, and black, slightly depressed spots are observed on stems and nuts. The blight can be controlled with timely and thorough applications of copper compounds.

A canker disease caused by the fungus *Sirococcus* sp. is causing a decline of butternut throughout much of its range. The canker does not appear to pose a serious threat to black walnut.

Phythophthora-induced root and crown rots have seriously curtailed seedling production of black and English walnut. Infected seedlings occur commonly in poorly drained portions of the nursery beds. The damage continues during winter storage. Some control has been obtained with fungicide drenches and dips. Species of *Cylindrocladium* also may be serious root rot pathogens of black walnut seedlings.

Not all defoliation of walnut trees is due to pathogenic organisms. The leaves on trees growing on poor sites or suffering from drought may turn yellow and drop during August or September as a direct result of environmental stress factors. Even under normal conditions, walnut trees drop their leaves earlier in the autumn than do many other species of trees. *See* PLANT PATHOLOGY. Dan Neely

Bibliography. R. A. Jaynes (ed.), *Nut Tree Culture in North America*, Northern Nut Growers Ass., 1979; D. Neely, R. Phares, and B. Weber, Cristulariella leaf spot associated with defoliation of black walnut plantations in Illinois, *Plant Dis. Rep.*, 60:587–590, 1976; B. Thompson, *Black Walnut for Profit*, 1976; U.S. Forest Service, Walnut insects and diseases, *Workshop Proceedings, June 13–14, 1978, Carbondale, Illinois*, 1979; J. G. Woodroof, *Tree Nuts*, 1979.

Warm-air heating system

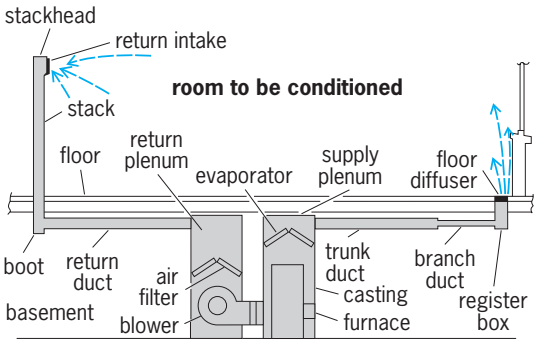
In a general sense, a heating system which circulates warm air. Under this definition both a room-heating stove and a steam blast coil circulate warm air. Strictly speaking, however, a warm-air system is one containing a direct-fired furnace surrounded by a bonnet through which air circulates to be heated (see **illus.**).

When air circulation is obtained by natural gravity action, the system is referred to as a gravity warm-air system. If positive air circulation is provided by means of a centrifugal fan (referred to in the industry as a blower), the system is referred to as a forced-air heating system.

Direct-fired furnaces are available for burning of solid, liquid, or gaseous fuels, although in recent years oil and gas fuels have been most commonly used. Furnaces have also been designed which have air circulating over electrical resistance heaters. A completely equipped furnace-blower package consists of furnace, burner, bonnet, blower, filter, and accessories. The furnace shell is usually of welded steel. The burner supplies a positively metered rate of fuel and a proportionate amount of air for combustion. A casing, or jacket, encloses the furnace and provides a passage for the air to be circulated over the heated furnace shell. The casing is insulated and contains openings to which return-air and warm-air ducts can be attached. The blower circulates air against static pressures, usually less than 1 in. (2.5 cm) water gage. The air filter removes dust particles from the circulating air. The most common type of air filter is composed of 1- to 2-in.-thick (2.5- to 5.0-cm) fibrous matting, although electrostatic precipitators are sometimes used. *See* GAS FURNACE; OIL BURNER.

Accessories to assure effective operation include automatic electrical controls for operation of burner and blower and safety control devices for protection against (1) faulty ignition of burner and (2) excessive air temperatures.

Ratings of warm-air furnaces are established from tests made in laboratories under industry-specified



Air passage in a warm-air duct system. (*After S. Konzo, J. R. Carroll, and H. D. Bareither, Winter Air Conditioning, Industrial Press, 1958*)

conditions. The tests commonly include heat-input rate, bonnet capacity, and register delivery. Heat-input rate is the heat released inside the furnace by the combustion of fuel, in Btu/h. Bonnet capacity refers to the heat transferred to the circulating air, in Btu/h. Register delivery is the estimated heat available at the registers in the room after allowance for heat loss from the ducts has been made, in Btu/h.

The recommended method for selection of a furnace is to estimate the total heat loss from the structure under design weather conditions, including the losses through the floor and from the basement, and to choose a furnace whose bonnet capacity rating is equal to, or greater than, the total design heat loss.

The complete forced-air heating system consists of the furnace-blower package unit; the return-air intake, or grille, together with return-air ducts leading from the grille to the return-air plenum chamber at the furnace; and the supply trunk duct and branch ducts leading to the registers located in the different spaces to be heated.

The forced-air system in recent years has no longer been confined to residential installations. The extreme flexibility of the system, as well as the diversity of furnace types, has resulted in widespread use of the forced-air furnace installations in the following types of installations, both domestic and commercial: residences with basement, crawl space, or with concrete floor slab; apartment buildings with individual furnaces for each apartment; churches with several furnaces for different zones of the building; commercial buildings with summer-winter arrangements; and industrial buildings with individual furnace-duct systems in each zone. *See* COMFORT HEATING.

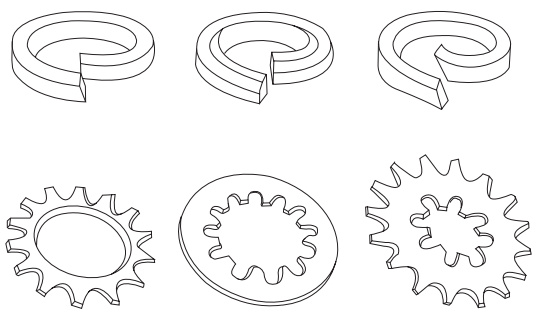
Seichi Konzo

Bibliography. American Society of Heating, Refrigerating, and Air-Conditioning Engineers, *Equipment*, 1992, *Fundamentals*, 1994; F. C. McQuiston and J. D. Parker, *Heating, Ventilating, and Air Conditioning: Analysis and Design*, 5th ed., 2000.

Washer

A flattened, ring-shaped device used to improve the tightness of a screw fastener. Three types of washer are in common use: plain, spring-lock, and antiturn (tooth-lock) washers. Standard plain washers are used to protect a part from damage or to provide for a wider distribution of the load. Because a plain washer will not prevent a nut from turning, a locking-type washer should be used to prevent a bolt or nut from loosening under vibration (see **illus.**). For industrial applications, spring-lock washers are intended to compensate for possible loosening between assembled parts and to facilitate assembly and disassembly. Lock washers create a continuous pressure between the parts and the fastener. The antiturn-type washers may be externally serrated, internally serrated, or both. The bent teeth bite into the bearing surface to prevent the nut from turning and the fastening from loosening under vibration. To speed up assembly, a variety of permanent preassembled bolt-and-washer and nut-and-washer combinations are available. *See* BOLT; NUT (ENGINEERING); SCREW FASTENER.

Warren J. Luzadder



Lock washers. (*After W. J. Luzadder, Fundamentals of Engineering Drawing, 6th ed., Prentice-Hall, 1971*)

Wastewater reuse

The use of treated wastewater effluent for beneficial purposes such as irrigation, industrial processes, and municipal water supply. Several developments have prompted wastewater reuse, including shortages of freshwater, stringent requirements for wastewater effluent quality, and advancements in treatment technology.

Shortage of freshwater. Freshwater is used for municipal, industrial, and irrigation purposes. Municipal and industrial water uses are nonconsumptive, whereas irrigation is consumptive, since most of this water is lost by evaporation and transpiration and is not available for reuse. Worldwide, the total cultivated area is about 1 billion hectares, of which only about one-third is irrigated. To increase food production, more irrigation water will be required, especially in developing countries.

The Earth's surface continually receives huge amounts of freshwater as precipitation. However, most precipitation is not available for human use due to its uneven distribution. The imbalance between the freshwater supply and demand has become more serious in recent years, with the available supply diminishing due to pollution and rapidly increasing demand. The World Health Organization estimates that within the next 50 years more than 40% of the world's population will live in countries facing water stress or water scarcity.

Requirements for wastewater effluent quality. A traditional sanitary engineering system deals with the municipal water supply and wastewater disposal (**Fig. 1**). Cities usually obtain their freshwater supply from surface or ground-water sources. After use, about 80% of the freshwater supply becomes wastewater, which is collected by a sewerage system and conveyed to a treatment plant. The wastewater is treated and then discharged as effluent into receiving water such as a river, estuary, coastal water, or ground-water aquifer.

Prior to 1970, most cities in the United States and other countries provided no more than primary wastewater treatment, which is a sedimentation

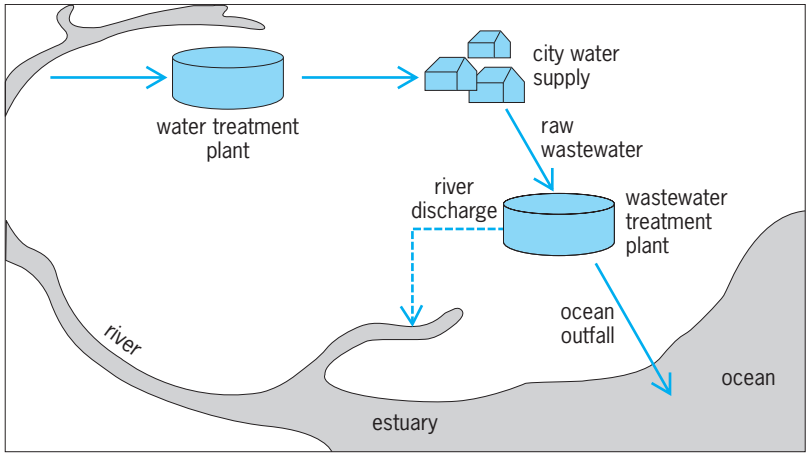


Fig. 1. Traditional water and wastewater engineering systems.

process that removes suspended solids. To restore and maintain the chemical, physical, and biological integrity of the nation's waters, Congress enacted the 1972 Federal Water Pollution Control Act Amendments (now called the Clean Water Act), which established nationwide minimum treatment requirements for all wastewater. For municipal wastewater discharge, the minimum is secondary treatment, which removes 85% of the biochemical oxygen demand (BOD) and total suspended solids. BOD is a measurement of oxygen-demanding organic wastes. Secondary treatment can be achieved by biochemical decomposition of organic waste materials, followed by further sedimentation. For industrial wastewater discharge, the minimum treatment is defined as the best practical control technology currently available. In situations in which these minimum treatment levels are not sufficient, the act requires additional treatment. *See* SEDIMENTATION (INDUSTRY); SEWAGE SOLIDS.

Water quality standards and required treatment processes. Wastewater reuse is classified as nonpotable and potable. Nonpotable uses of wastewater effluent include irrigation, industrial process or cooling water, and recreational impoundments (such as lakes or reservoirs for public use). Potable uses of high-quality wastewater effluent can be achieved by sending it back to water supply reservoirs, to potable ground-water aquifers, or directly to water treatment plants.

There are no federal regulations governing wastewater reuse in the United States. All regulations are enforced at the state level. As the U.S. Clean Water Act requires, municipal wastewater must receive at least secondary treatment; therefore, the reuse potential of wastewater depends on levels of postsecondary (or advanced/tertiary) treatment imposed. The Hawaii Department of Health has guidelines for three types of reclaimed water and the treatment and limitations of each. The highest-quality reclaimed water, designated R-1 water, is wastewater that has received secondary treatment followed by filtration and intense disinfection. R-1 water is deemed acceptable for public contact and can be used for nearly all nonpotable purposes. Next in quality is R-2 water, which is wastewater that has received secondary treatment followed by disinfection. The lowest-quality reclaimed water, designated R-3 water, is wastewater that has received secondary treatment but no disinfection. The reuse of R-2 and R-3 waters is restricted. *See* FILTRATION; SEWAGE TREATMENT.

Fecal coliform bacteria found in the intestinal tract of warm-blooded animals are used as an indicator of microbiological purity. The amount of fecal coliform in a water sample is measured in colony-forming units (CFU). Most states set a maximum of 23 CFU/100 milliliters for unrestricted irrigation use or irrigation of edible crops, sports fields, and public parks. California and Arizona set a maximum of 2.2 CFU/100 ml for higher-quality reclaimed water that can be used for recreational impoundment.

Artificial ground-water recharge of reclaimed water into potable aquifers can be accomplished by using injection wells or by surface spreading. The reclaimed water must meet drinking water standards if it is recharged into potable aquifers, and it must meet drinking water standards after percolation through the vadose zone if it is recharged by surface spreading. The vadose zone (or unsaturated zone) occurs immediately below the land surface, where soil pores (or interconnected openings) may contain either air or water.

Heavy metals and other trace toxic substances can enter municipal wastewater systems that receive industrial discharge. In these cases, special monitoring and advanced treatment of wastewater effluent before reuse are necessary. Advanced treatment methods, such as carbon adsorption, precipitation, and pressure-driven membrane processes, are used to further remove dissolved solids or salts, nutrients, and organic and inorganic chemicals. *See* ACTIVATED CARBON; PRECIPITATION (CHEMISTRY); WATER TREATMENT.

Pressure-driven membrane technology. Pressure-driven membranes are divided into four groups based on membrane pore size: microfiltration (MF), ultrafiltration (UF), nanofiltration (NF), and reverse osmosis (RO). The required pressure for each membrane type is inversely proportional to the pore size. Microfiltration is used to remove the largest particles and requires the least pressure, whereas the reverse osmosis is used to remove the smallest particles and requires the greatest pressure. Microfiltration and ultrafiltration use a sieve process, in which particles are removed because the membrane openings are smaller than the particles' size. When two solutions of salts are separated by a semipermeable membrane, osmosis allows water molecules to pass through the membrane from the solution of lower salt concentration into the solution of higher salt concentration. The reverse osmosis process, which reverses the direction of water, can be accomplished by applying to the higher-salt-concentration solution a pressure that is higher than the natural osmotic pressure. Nanofiltraton uses the sieve and reverse osmosis processes. *See* MEMBRANE SEPARATIONS; OSMOSIS; ULTRAFILTRATION.

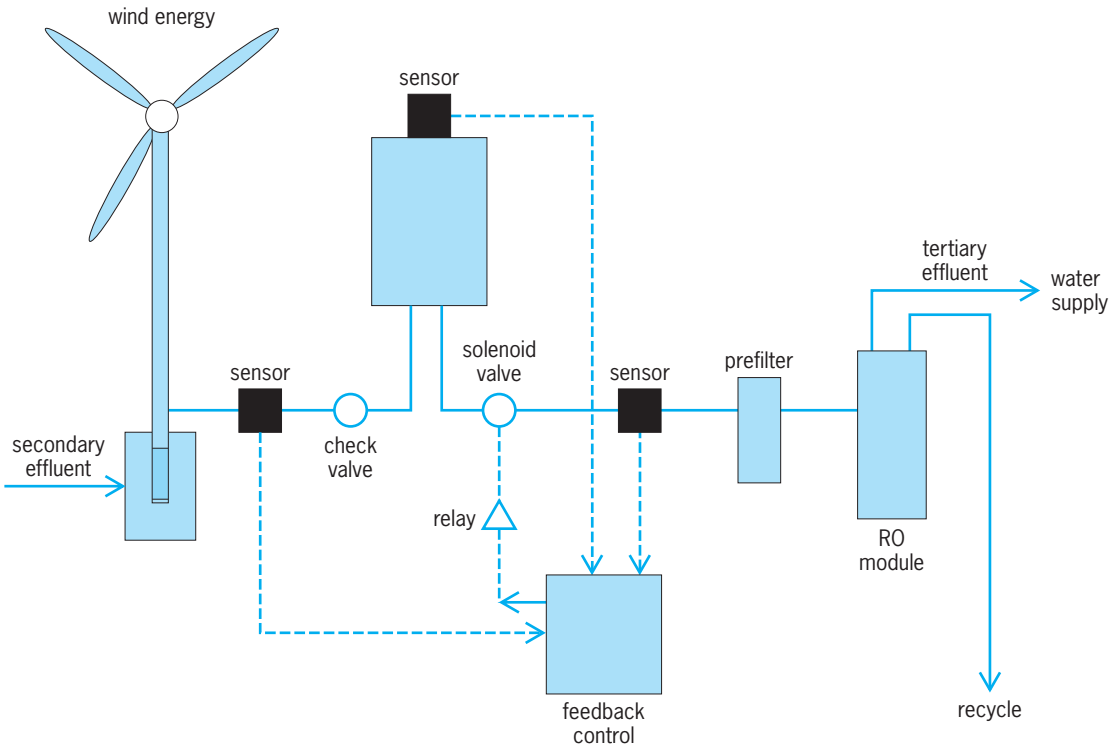


Fig. 2. Wastewater-reuse system with a wind-driven reverse osmosis process.

Reverse osmosis was developed in the 1960s with the advent of the asymmetric cellulose acetate membrane. Since then, its popularity has steadily risen because of advances made in membrane technology with the use of different materials and membrane configurations. Progress has also been made in overcoming membrane fouling and extending the life of the membrane.

Pressure-driven membrane technology is the key to expanding wastewater reuse in the future. Pressure-driven membranes can provide effective disinfection as well as total dissolved solids (TDS) removal. Pressure-driven membranes are ideal for postsecondary treatment of municipal wastewater in coastal areas. In these areas, brackish water or seawater enters the sewer infrastructure due to infiltration and inflow, causing high salinity or TDS in the wastewater flow into treatment plants. Recently, a new treatment technology called membrane bioreactor (MBR) was developed by combining micro- or ultrafiltration membranes with secondary or biological wastewater treatment processes. The city of San Diego is developing a plan to use reverse osmosis, along with a few pretreatment and posttreatment processes, to reclaim and purify its wastewater effluent before returning it to the city's water supply reservoir. Water management authorities in the Netherlands are also studying the use of membrane technology in wastewater treatment.

The high amount of energy required to create and maintain feed water pressure is a limiting factor for expanding the use of pressure-driven membranes in water and wastewater treatment systems. To address this problem, new technology is being developed that integrates the use of natural energy with pressure-driven membranes. A prototype wastewater reuse system with a wind-driven reverse osmosis process was constructed by University of Hawaii researchers on Coconut Island off the windward coast of Oahu. This system consists of a 30-ft-tall (10-m) multivaned windmill, an ultra-low-pressure RO membrane, a flow/pressure stabilizer, a prefilter, and a feedback control mechanism (**Fig. 2**). Under a moderate wind speed of 11 mi/h (5 m/s), this system can process a flow of 5000 gallons/day (19 m^3/day) and reduce the TDS content from 3100 mg/L to 50 mg/L. Another new technology is ocean-wave-driven RO desalination, which was successfully tested in the Caribbean islands by University of Delaware researchers. *See* WATER DESALINATION; WIND POWER.

As the shortage of freshwater becomes more serious and the quality of wastewater effluent improves, reuse of municipal wastewater will become one of the most attractive water resource management alternatives. Advances in health risk assessment will make treated wastewater a widely acceptable freshwater source. The field of water and wastewater engineering is undergoing rapid changes. It is expected that, through proper source control and treatment, water and wastewater will be indistinguishable. *See* WATER SUPPLY ENGINEERING.

Clark C. K. Liu

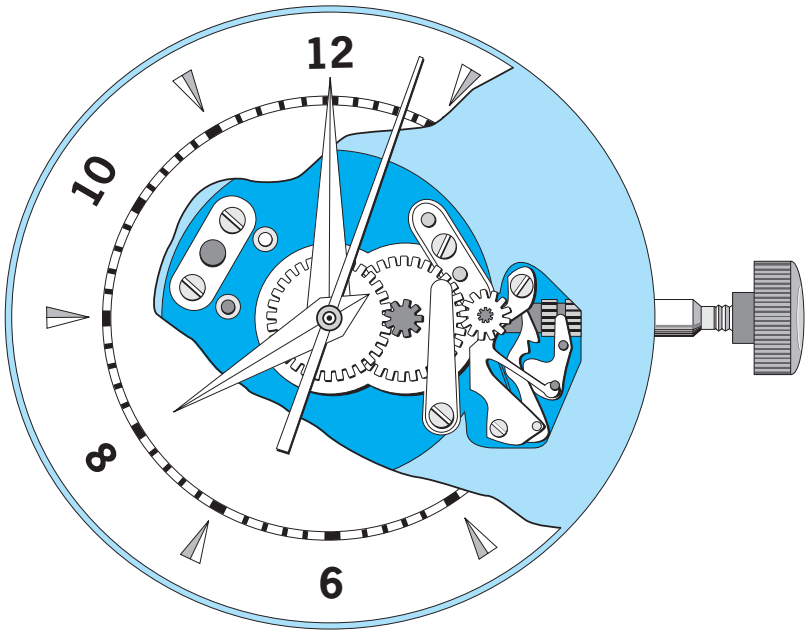
Bibliography. R. Aertgeerts and A. Angelakis (eds.), *State of the Art Report Health Risks in Aquifer Recharge Using Reclaimed Water*, World Health Organization, Regional Office for Europe, Copenhagen, 2003; H. Bouwer, Role of groundwater recharge in treatment and storage of wastewater

for reuse, *Water Sci. Technol.*, 24(9):295–302, 1991; Y. Kamiyama et al., New thin-film composite reverse osmosis membranes and spiral wound modules, *Desalination*, 51:79–92, 1984; T. Richardson and R. Trussel, Taking the plunge, *Civ. Eng.*, 67(9):42–45, 1997; D. Seckler et al., *World Water Demand and Supply, 1990 to 2025: Scenarios and Issues*, Res. Rep. 19, International Water Management Institute, Colombo, Sri Lanka, 1998; U.S. Environmental Protection Agency, *Municipal Wastewater Reuse: Selected Readings on Water Reuse*, EPA430/09-91-002, Washington, DC, 1991.

Watch

A portable instrument that measures time. The direct forerunner of the watch was the pendulum clock, invented by Christian Huygens in 1656.

Principles of operation. The creation of the pendulum clock began with the insights of Galileo in the fifteenth century. Galileo, attending services at the Cathedral of Pisa, noticed that when a chandelier was caused to swing back and forth the long or short swings were completed in the same amount of time. He knew that this was because during the long swing the increased gravitational pull made the chandelier swing faster and during the short swing the decreased gravitational pull made the chandelier swing slower. The result was that the swings—long or short—took the same amount of time. Galileo verified this by his timing apparatus: his pulse. Christian Huygens used this principal "of equal time for long or short swing" to invent the first pendulum clock in 1656. For this clock, power was supplied not by gravity but by a wound spring; whether the spring is tightly or loosely wound, the pendulum (like the chandelier) swings in equal time. *See* PENDULUM.



Cutaway view of classic mechanical watch.

The principles discovered by Galileo and Huygens are the precursors of all modern timepieces, which operate by using the following components: a part that swings back and forth in equal intervals called an oscillator (for example, a pendulum, balance wheel, tuning fork, or quartz crystal); a power source that energizes the swings (for example, a mainspring, battery, or capacitors); and a system that counts the swings (see **illustration**). The earliest watches consisted of a "swinger" (basically a balance wheel with an attached hairspring) as an oscillator; a mainspring as a power source; and a dial and hands to count the swings. This type of model is still manufactured and is comparatively expensive. Some are called chronometers and are precise timekeepers. But most watches today are quartz watches. *See* CLOCK; TUNING FORK.

Evolution of the quartz watch. Quartz watches are the result of the evolution of the clock mechanism. The first step was to replace the mechanical power (the mainspring) with electrical power (the battery). The next major step was to replace the oscillator (the balance wheel, with its attached hairspring) with a tuning fork and an electronic circuit; although this mechanism was eventually discontinued, the electronic circuit was retained in later watches. The next step, which was truly revolutionary, was to replace the tuning fork with a quartz crystal. Whereas a pendulum swings once per second, and a balance wheel oscillates 5 times a second, the standard quartz crystal oscillates 32,768 times a second.

Quartz crystals. The advent of the quartz crystal is the most revolutionary advance in watchmaking in the twentieth century. The properties of the quartz crystal were discovered by Pierre and Jacques Curie. They experimented with natural crystals and discovered piezoelectricity (the principle that when crystals are stressed they generate electricity). The most responsive was the quartz crystal. Since the quartz crystal produced extremely stable oscillations, it was applied to frequency control units in radio broadcasting. Then, in 1928, Bell Laboratories created the first quartz crystal clock. But before a quartz watch could be produced an integrated circuit had to be created. *See* PIEZOELECTRICITY; QUARTZ CLOCK.

Integrated circuits. The integrated circuit, or microchip, is one of the most revolutionary inventions of the twentieth century. It is used in every type of electronic device—from computers to antimissile systems. In the past, electronic circuits consisted of vacuum tubes, diodes, transistors, capacitors, and resistors, which were connected by wires and fastened to a baseboard. Depending on the number of elements involved, such a circuit could require an entire room or building. By the process of miniaturization the electronic circuit was transferred into the integrated circuit (microchip). In a watch, the microchip reduces the frequency of the quartz vibrations (32,768 times per second) to standard intervals. The microchip was patented in 1959, and in 1967 Japan and Switzerland simultaneously introduced quartz watches at the Basel Fair. *See* INTEGRATED CIRCUITS.

Stepping motors. In a quartz watch the quartz crystal requires an intermediary to reduce the oscillations to activate the setting system for dial and hands. This is accomplished with a stepping motor, which is energized by the integrated circuit (microchip). In the earliest quartz watch models "slave" oscillators (for example, balance wheels or tuning forks), activated by the microchip, were used to reduce the oscillations. However, eventually all manufactures began to use stepping motors. *See* STEPPING MOTOR.

Solid-state technology. After the quartz watch with its stepping motor was invented another historic development took place—the solid-state quartz, a watch that displayed the time with numbers (digits) rather than with hands and a dial. Whereas the hands on the dial of a stepping-motor quartz watch are moved by a set of gears attached to a setting mechanism, and an integrated circuit divides the quartz oscillations to one pulse per second (or fraction thereof), in a solid-state watch the time display is formed strictly electronically and inert matter forms the digits on the display screen. The integrated circuit controls the voltages and currents that activate the display.

With solid-state technology the revolution from a mechanical movement to a quartz movement was finally completed. No trace of the original mechanical movement remains in quartz watches, which rely solely upon electronic impulses. Freed from moving parts the full potential of the integrated circuit was realized. The integrated circuit in the solid-state watch is actually a miniature computer.

Liquid-crystal technology. An important advance in solid-state watches resulted from the discovery that liquid crystals could be rearranged and form configurations by electrical excitation. Light-emitting diodes (LEDs) were replaced by liquid crystal displays (LCDs), which are now used in all digital-reading solid-state quartz watches. *See* ELECTRONIC DISPLAY; LIGHT-EMITTING DIODE; LIQUID CRYSTALS.

Further advances. Since the quartz watch is also a miniaturized computer, it may contain many additional features. For example, some watches can import information such as appointments, birthdays, and addresses from a personal computer and display them on their screen. For the vision-impaired, watches can be made to "talk." For the outdoor sportsman, a watch exists that provides readouts of height, depth, and barometric pressure measurements. Some watches can display caloric consumption, based on exercise, if programmed with the wearer's age, weight, height, and sex; and there is a watch in which blood pressure and pulse rates appear on the screen and which flashes a warning across the screen if the wearer's blood pressure is too high. The face of a watch, whether analog (dials and hands) or digital, can in some cases be made to light up in the dark, a feature which has saved lives in emergency situations.

One notable advance in quartz watch technology started with a watch that picked up time signals from an astronomical observatory and thereby corrected the oscillations of the quartz crystal. This was accomplished by adding a miniature receiving set and antennae to the watch. This led to the development of a watch that is used as an entrance opener in over 300 resorts worldwide—it contains a computer chip which stores the code numbers of the entrances and an antenna which transmits the code to the entrance gate.

There are watches that have a camera attached. When a button is pressed the display screen becomes a viewfinder, and another button snaps the picture; the pictures can then be transferred to a computer. There are watches that receive and play music from the Internet or a compact disk; phone watches that contain a built-in microphone for outgoing calls and a built-in speaker for incoming calls; and a TV/phone watch that can receive television programs and also contains a phone (if a call comes in when the TV is on the watch automatically turns on the phone function). The advances possible in watch technology are as boundless as the capabilities of computer technology. *See* DIGITAL COMPUTER; TELEPHONE.

Benjamin Matz

Bibliography. T. Hyltin, *The Digital Electronic Watch*, 1978; B. Matz, *The History and Development of the Quartz Watch*, 1999; T. R. Reid, *The Chip: How Two Americans Invented the Microchip and Launched a Revolution*, 1984; H. Tait, *Clocks and Watches*, 1983.

Water

The chemical compound with two atoms of hydrogen and one atom of oxygen in each of its molecules. It is formed by the direct reaction (1) of hydrogen

$$2H_2 + O_2 \rightarrow 2H_2O \quad (1)$$

with oxygen. The other compound of hydrogen and oxygen, hydrogen peroxide, readily decomposes to form water, reaction (2). Water also is formed in the

$$2H_2O_2 \rightarrow 2H_2O + O_2 \quad (2)$$

combustion of hydrogen-containing compounds, in the pyrolysis of hydrates, and in animal metabolism. Some properties of water are given in the **table**.

Gaseous state. Water vapor consists of water molecules which move nearly independently of each other. The relative positions of the atoms in a water

Properties of water

Property	Value
Freezing point	0°C
Density of ice, 0°C	0.92 g/cm^3
Density of water, 0°C	1.00 g/cm^3
Heat of fusion	80 cal/g (335 J/g)
Boiling point	100°C
Heat of vaporization	540 cal/g (2260 J/g)
Critical temperature	347°C
Critical pressure	217 atm (22.0 MPa)
Specific electrical conductivity at 25°C	1×10^{-7}/ohm-cm
Dielectric constant, 25°C	78

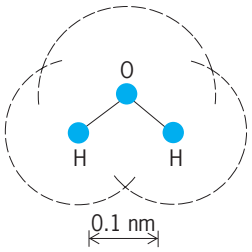


Fig. 1. Water molecule.

molecule are shown in **Fig. 1**. The dotted circles show the effective sizes of the isolated atoms. The atoms are held together in the molecule by chemical bonds which are very polar, the hydrogen end of each bond being electrically positive relative to the oxygen. When two molecules near each other are suitably oriented, the positive hydrogen of one molecule attracts the negative oxygen of the other, and while in this orientation, the repulsion of the like charges is comparatively small. The net attraction is strong enough to hold the molecules together in many circumstances and is called a hydrogen bond. *See* CHEMICAL BONDING; ELECTRONEGATIVITY; GAS; VALENCE.

When heated above 1200°C (2200°F), water vapor dissociates appreciably to form hydrogen atoms and hydroxyl free radicals, reaction (3). These products recombine completely to form water when the temperature is lowered. Water vapor also undergoes most of the chemical reactions of liquid water and, at very high concentrations, even shows some of the unusual solvent properties of liquid water. Above 374°C (705°F), water vapor may be compressed to any density without liquefying, and at a density as high as 0.4 g/cm^3, it can dissolve appreciable quantities of salt. These conditions of high temperature and pressure are found in efficient steam power plants. *See* HYDROGEN BOND.

$$H_2O \rightarrow H + OH \qquad (3)$$

Solid state. Ordinary ice consists of water molecules joined together by hydrogen bonds in a regular arrangement, as shown in **Fig. 2**. The circles represent only the positions of the atoms, but if the sizes, as indicated in Fig. 1, are superimposed upon the figure, then it appears that there is considerable empty space between the molecules. This unusual feature is a result of the strong and directional hydrogen bonds taking precedence over all other intermolecular forces in determining the structure of the crystal. If the water molecules were rearranged to reduce the amount of empty space, their relative orientations would no longer be so well suited for hydrogen bonds. This rearrangement can be produced by compressing ice to pressures in excess of 2000 atm (14 megapascals). Altogether five different crystalline forms of solid water have been produced in this way, the form obtained depending upon the final pressure and temperature. They are all more dense than water, and all revert to ordinary ice when the pressure is reduced. *See* CRYSTAL STRUCTURE.

Liquid state. The molecules in liquid water also are held together by hydrogen bonds. When ice melts, many of the hydrogen bonds are broken, and those that remain are not numerous enough to keep the molecules in a regular arrangement. Many of the unusual properties of liquid water may be understood in terms of the hydrogen bonds which remain. As water is heated from 0°C (32°F), it contracts until 4°C (39°F) is reached and then begins the expansion which is normally associated with increasing temperature. This phenomenon and the increase in density when ice melts both result from a breaking down of the open, hydrogen-bonded structure as the temperature is raised. The viscosity of water decreases tenfold as the temperature is raised from 0 to 100°C (32 to 212°F), and this also is associated with the decrease of icelike character in the water as the hydrogen bonds are disrupted by increasing thermal agitation. Even at 100°C (212°F), the hydrogen bonds influence the properties of water strongly, for it has a high boiling point and a high heat of vaporization compared with other substances of similar molecular weight. *See* LIQUID.

The electrical conductivity of water is at least 1,000,000 times larger than that of most other nonmetallic liquids at room temperature. The current in this case is carried by ions produced by the dissociation of water according to reaction (4). This reaction is reversible, and equilibrium is reached rapidly, so there is a definite concentration of H^+ and OH^- ions in pure water. At 25°C (77°F), this concentration is 10^{-7} mole/liter of each species or about 10^{14} ions/ml. This concentration of ions is affected by the temperature or by the presence of solutes in the water. *See* ACID AND BASE; HYDROGEN ION.

$$H_2O \rightleftharpoons H^+ + OH^- \qquad (4)$$

Pure water, either solid or liquid, is blue if viewed through a thickness of more than 6 ft (2 m). The other colors often observed are due to impurities.

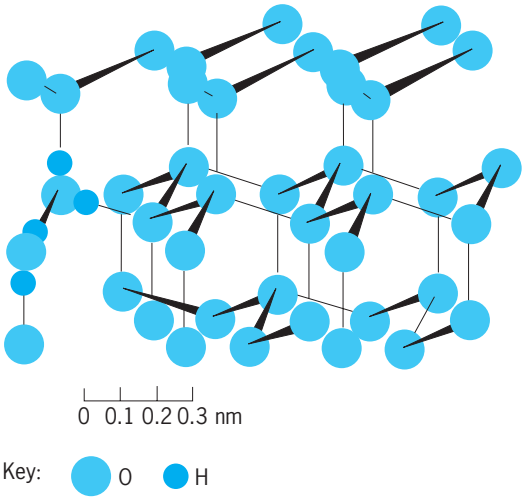


Fig. 2. Structure of ice. The hydrogen atoms are omitted for all but two water molecules.

Solutions in water. Water is an excellent solvent for many substances, but particularly for those which dissociate to form ions. Its principal scientific and industrial use as a solvent is to furnish a medium for purifying such substances and for carrying out reactions between them. *See* SOLUTION; SOLVENT.

Among the substances which dissolve in water with little or no ionization and which are very soluble are ethanol and ammonia. These are examples of molecules which are able to form hydrogen bonds with water molecules, although, except for the hydrogen of the OH group in ethanol, it is the hydrogen of the water that makes the hydrogen bond. On the other hand, substances which cannot interact strongly with water, either by ionization or by hydrogen bonding, are only sparingly soluble in it. Examples of such substances are benzene, mercury, and phosphorus. For discussions of another important class of solutions in water *see* COLLOID; SURFACTANT.

Chemical properties. Water is not a strong oxidizing agent, although it may enhance the oxidizing action of other oxidizing agents, notably oxygen. Examples of the oxidizing action of water itself are its reactions with the alkali and alkaline-earth metals, even in the cold; for instance reaction (5), and its reactions with iron and carbon at elevated temperatures, reactions (6) and (7). Reaction (7) is used commercially to

$$Ca + 2H_2O \rightarrow Ca^{2+} + 2OH^- + H_2 \quad (5)$$

$$3Fe + 4H_2O \rightarrow Fe_3O_4 + 4H_2 \quad (6)$$

$$C + H_2O \rightarrow CO + H_2 \quad (7)$$

produce a gaseous fuel from solid coke. The gaseous mixture, $CO + H_2$, called water gas, is formed when steam is passed over coke heated to 600°C (1100°F).

Water is an even poorer reducing agent than oxidizing agent. One of the few substances that it reduces rapidly is fluorine, but this reaction is complicated. Chlorine is reduced only very slowly in the cold, according to reaction (8).

$$2Cl_2 + 2H_2O \rightarrow O_2 + 4H^+ + 4Cl^- \quad (8)$$

An example of another sort of oxidation-reduction reaction in which water plays an essential role beyond that of the solvent is the disproportionation of chlorine, reaction (9), which is fast and incomplete

$$Cl_2 + H_2O \rightarrow HOCl + H^+ + Cl^- \quad (9)$$

in neutral solution but goes to completion if base is added. *See* OXIDATION-REDUCTION.

Substances with strong acidic or basic character react with water. For example, calcium oxide, a basic oxide, reacts in a process called the slaking of lime, reaction (10). Sulfur trioxide, an acidic oxide, also reacts, reaction (11). This reaction occurs in the contact process for the manufacture of sulfuric acid. Both of these reactions evolve enough heat to produce fires or explosions unless precautions are taken.

$$CaO + H_2O \rightarrow Ca(OH)_2 \quad (10)$$

$$SO_3 + H_2O \rightarrow H_2SO_4 \quad (11)$$

Another type of substance with strong acidic character is an acid chloride. Two examples and their reactions with water are boron trichloride, reaction (12), and acetyl chloride, reaction (13). These are

$$BCl_3 + 3H_2O \rightarrow H_3BO_3 + 3HCl \quad (12)$$

$$CH_3COCl + H_2O \rightarrow CH_3CO_2H + HCl \quad (13)$$

often termed hydrolysis reactions, as in the reaction of an ester with water, for instance ethyl acetate, reaction (14). A hydrolysis reaction of a different sort is that of calcium carbide, used in the production of acetylene, reaction (15).

$$H_3CCOC_2H_5 + H_2O \longrightarrow H_3CCOH + HOC_2H_5 \quad (14)$$

$$CaC_2 + 2H_2O \rightarrow Ca(OH)_2 + C_2H_2 \quad (15)$$

See HYDROLYSIS.

Water reacts with a variety of substances to form solid compounds in which the water molecule is intact, but in which it becomes a part of the structure of the solid. Such compounds are called hydrates, and are formed frequently with the evolution of considerable amounts of heat. Examples range from the hydrates of simple and double salts, calcium chloride hexahydrate, $CaCl_2 \cdot 6H_2O$, and ammonium aluminum alum, $NH_4Al(SO_4)_2 \cdot 12H_2O$, to the gas hydrates which are stable only at low temperatures, for example, chlorine hydrate, $Cl_2 \cdot 6H_2O$, and xenon hydrate, $Xe \cdot 6H_2O$. *See* CLATHRATE COMPOUNDS; HYDRATE.

For various aspects of water, its uses, and occurrence *see* HEAVY WATER; HYDROGEN; HYDROLOGY; IRRIGATION (AGRICULTURE); OXYGEN; PLANT-WATER RELATIONS; PRECIPITATION (METEOROLOGY); SEAWATER; TRIPLE POINT; VAPOR PRESSURE; WATER POLLUTION; WATER SOFTENING; WATER SUPPLY ENGINEERING; WATER TABLE; WATER TREATMENT; WATERPOWER.

Harold L. Friedman

Water-borne disease

Disease acquired by drinking water contaminated at its source or in the distribution system, or by direct contact with environmental and recreational waters. Water-borne disease results from infection with pathogenic microorganisms or chemical poisoning.

These pathogenic microorganisms include viruses, bacteria, protozoans, and helminths. A number of microbial pathogens transmitted by the fecal-oral route are commonly acquired from water in developing countries where sanitation is poor. Viral pathogens transmitted via fecally contaminated water include hepatitis viruses A and E. Important bacterial pathogens transmitted via fecally

contaminated water in the developing world are *Vibrio cholerae*, enterotoxigenic *Escherichia coli*, *Shigella*, and *Salmonella enterica* serotype Typhi. Water-borne protozoan pathogens in the developing world include *Giardia lamblia* and *Entamoeba histolytica*. The major water-borne helminthic infection is schistosomiasis; however, transmission is not fecal-oral. Eggs of the trematode *Schistosoma* hatch in fresh water, and the larvae invade specific snail hosts. Snails release *Schistosoma* forms called cercariae which penetrate the skin when humans contact infested waters. Eggs are excreted from infected humans in feces or urine, depending on the *Schistosoma* species. Another water-borne helmenthic infection is dracunculiasis (guinea worm infection). The nematode, *Dracunculus medinensis*, is acquired by ingestion of water containing copepods infested with larvae infectious for humans. Larvae infectious for copepods are released by infected humans into water through skin ulcers caused by the adult worm. Dracunculiasis has been the object of an effective global eradication program involving water filtration, elimination of the copepod vector, and education of populations at risk.

In developed countries, fecal contamination of drinking water supplies is less likely. However, there have been outbreaks of diseases such as shigellosis and giardiasis associated with lapses in proper water treatment, such as cross-contamination of waste-water systems and potable water supplies. Animals are therefore more likely to play a role in water-borne disease in developed countries. Bacterial pathogens acquired from animal feces such as nontyphoid *S. enterica*, *Campylobacter jejuni*, and *E. coli* serotype O157:H7 have caused outbreaks of water-borne disease in developed countries where water is not properly chlorinated. Hikers frequently acquire *G. lamblia* infections from drinking untreated lake and stream water. *Giardia lamblia* may have animal reservoirs and can persist in the environment. A recently recognized pathogen apparently resistant to standard chlorination and filtration practices is the protozoan *Cryptosporidium parvum*. This organism is found in the feces of farm animals and may enter water supplies through agricultural runoff. Very large urban outbreaks of diarrheal disease acquired through municipal water supplies have been caused by this pathogen. Filtration reduces the risk of infection but does not completely eliminate this pathogen, for which the infectious dose for humans appears to be very small. A marine dinoflagellate pathogenic for fish, *Pfiesteria piscicida*, has recently been implicated on the Atlantic coast of the United States in disease syndromes involving memory loss. However, the studies are not yet conclusive.

Chemical poisoning of drinking water supplies causes disease in both developing and developed countries. Lead, copper, and cadmium have been frequently involved. *See* CHOLERA; ESCHERICHIA; MEDICAL PARASITOLOGY; SCHISTOSOMIASIS.

Steve L. Moseley

Bibliography. A. S. Benenson, *Control of Communicable Diseases in Man*, 16th ed., American Public Health Association Publications, 1995.

Water conservation

The protection, development, and efficient management of water resources for beneficial purposes. Water's physical and chemical properties make it essential to life and civilization. It is combined with carbon dioxide by green plants in the synthesis of carbohydrates, from which all other foods are formed. It is a highly efficient medium for dissolving and transporting nutrients through the soil and throughout the bodies of plants and animals. It can also carry deadly organisms and toxic wastes. Water is a raw material that is indispensable for many domestic and industrial purposes.

Water Distribution and Cycles

Water occupies more than 71% of the Earth's surface. Over 97% of the (1.386×10^9 km^3) of water is in oceans, and a little over 2.5% is fresh water. Icecaps, glaciers, and other forms of permanent snow that cover the Antarctic, Arctic, and mountainous regions make up 77% of the fresh water; ground water is 22% of all fresh water; and the other 1% of fresh water is divided among the lakes (61%), and atmosphere and soil moisture (39%), and streams (less than 0.4%). Although streams (which include rivers) constitute the smallest percentage of the water distribution, they provide the majority of water used by people.

Hydrologic cycle. The hydrologic cycle is the continuous circulation of water from the ocean to the atmosphere to the land and back to the ocean (**Fig. 1**). Water is temporarily stored in streams, lakes, soil, ice, and underground before it becomes available for use. The ocean stores most of the water, which reenters the cycle when it evaporates due to solar energy. Water is carried through the atmosphere, becoming precipitation (in the form of rain or snow) under the right conditions.

Precipitation. Precipitation occurs when warm air masses that carry moisture are cooled, which decreases their capacity to retain moisture to the point at which water is released.

The source of the air masses and the pressure and temperature gradients aloft determine the form, intensity, and duration of precipitation. Precipitation may occur as cyclonic or low-pressure storms (mostly a winter phenomenon responsible for widespread rains), as thunderstorms of high intensity and limited area, or as mountain storms wherein warm air dumps its moisture when lifted and cooled in crossing high land barriers. *See* PRECIPITATION (METEOROLOGY).

Precipitation is characteristically irregular. Generally, the more humid an area and the nearer the ocean, the more evenly distributed is the rainfall. Whatever the annual average, rainfall in arid regions

tends to vary widely from year to year and to fall in a few heavy downpours. Large floods and active erosion result from heavy and prolonged rainfall or rapid melt of large volumes of snow. Flash floods often follow local intense thunderstorms.

Storage. As rain falls to the Earth, it is caught by vegetation. Although some water is stored on the vegetation itself, most of it runs off onto the ground. The small portion that is stored on the vegetation is called interception. Interception results in evaporation of the water back into the atmosphere. Water that does reach the ground can be absorbed by the soil, remain on the surface and run off into lakes and streams, evaporate from the ground surface back into the atmosphere, or be utilized by vegetation for photosynthesis or transpiration. *See* PLANT-WATER RELATIONS.

Water that seeps into the soil and is held in place as soil moisture by capillary forces is called infiltration. Water that moves over the ground surface and runs off into lakes and streams is called overland flow. Older soil water will eventually become displaced with infiltrating water as soil moisture increases with increasing precipitation, and will result in subsurface runoff. Subsurface runoff can move laterally through the soil or vertically into the ground-water zone, where soil pores and rock are completely filled with water. Ground water will slowly move into streams, lakes, and wetlands and provide surface water.

Recharge. The full recharge (replenishment) of water sources varies according to each water storage element. For example, the full recharge of permafrost is approximately 10,000 years, and polar ice occurs every 9700 years. Alpine glacial recharge is around 1600 years. Recharge of oceanic waters occurs about 2500 years, and ground-water recharge can average 1400 years. Recharge of lakes, bogs, and wetlands ranges between 5 and 17 years, whereas recharge of the atmosphere and stream channel networks ranges between 8 and 16 days.

The hydrologic cycle is a critical framework for identifying how human activity modifies land and water resources. Human activities alter land surface conditions, how water is stored in various parts of the hydrologic cycle, how quickly that storage can be recharged, and how water is distributed across the landscape. Understanding what aspects of the hydrologic cycle are affected by human activities can help to target conservation measures more effectively.

Underground water. Usable ground water in the United States is estimated to be 4.75×10^{10} acre-feet. Annual runoff from the land averages 1.299×10^{9} acre-feet (1.602×10^{12} m^3). The volume of ground water greatly exceeds that of all fresh-water lakes and reservoirs combined. It occurs in several geologic formations (aquifers) and at various depths. Ground water under pressure is known as artesian water, and it may become available either by natural or artificial flowing wells. Ground water, if abundant, may maintain streams and springs during extended dry periods. It originates from precipitation of various ages as determined by measurements

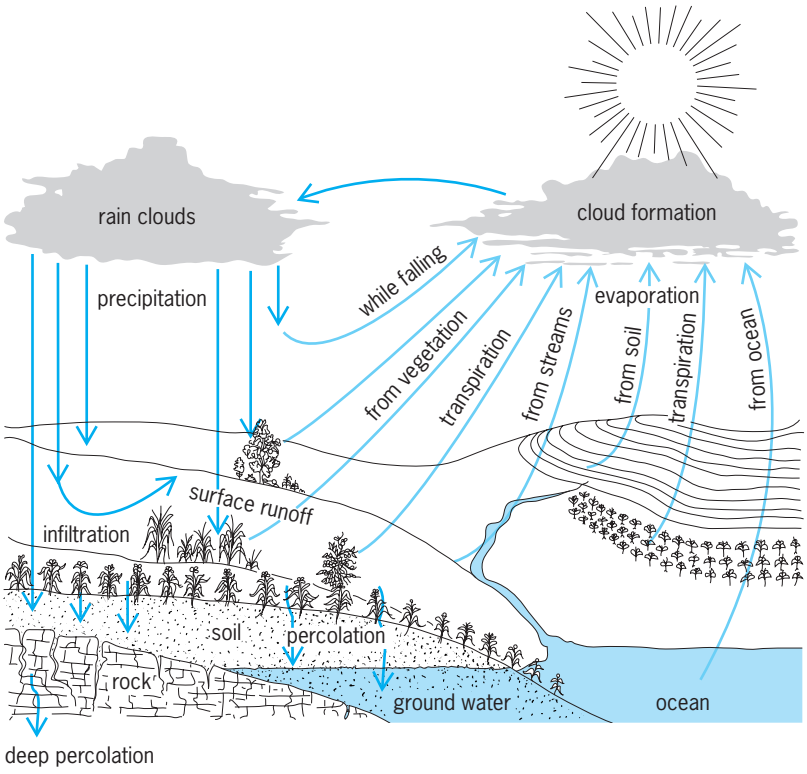


Fig. 1. Hydrologic cycle. (*Water, USDA Yearbook, 1955*)

of the decay of tritium, a radioisotope of hydrogen found in ground water. The water table is the upper level of saturated ground-water accumulation. It may appear at the surface of the Earth in marshes, swamps, lakes, or streams, or hundreds of feet down. Seeps or springs occur where the contour of the ground intercepts the water table. In seeps the water oozes out, whereas springs have distinct flow. Water tables fluctuate according to the source and extent of recharge areas, the amount and distribution of rainfall, and the rate of extraction. The yield of aquifers depends on the porosity of their materials. The yield represents that portion of water which drains out by gravity and becomes available by pumping. Shallow ground water (down to 50 ft or 15 m) is trapped by dug or driven wells, but deep sources require drilled wells. The volume of shallow wells may vary greatly in accordance with fluctuations in rainfall and degree of withdrawal. *See* GROUND-WATER HYDROLOGY.

Surface water. Streams supply most of the water needs of the United States. Lakes, ponds, swamps, and marshes, like reservoirs, represent stored streamflow. The natural lakes in the United States are calculated to contain 1.3×10^{10} acre-feet (1.6×10^{13} m^3). Swamps and other wetlands along river deltas, around the borders of interior lakes, and in coastal regions add millions more to the surface supplies. The oceans and salty or brackish sounds, bays, bayous, or estuaries represent almost unlimited potential fresh-water sources. Brackish waters are being used increasingly by industry for cooling and flushing. Reservoirs, dammed lakes, farm ponds, and other small impoundments have a combined usable

storage of 3×10^8 acre-feet (3.7×10^{11} m^3). The smaller ones furnish water for livestock, irrigation, fire protection, flash-flood protection, fish and waterfowl, and recreation. However, most artificial storage is in reservoirs of over 5000 acre-feet (6.165×10^6 m^3). Lake Mead, located in Arizona and Nevada and formed by Hoover Dam, is the largest (227 mi^2 or 588 km^2) of the 1300 reservoirs, and it contains 10% of the total stored-water capacity, or over 3.1×10^7 acre-feet (3.8×10^{10} m^3). These structures regulate streamflow to provide more dependable supplies during dry periods when natural runoff is low and demands are high. They store excess waters in wet periods, thus mitigating damaging floods. *See* DAM; SURFACE WATER.

Human Impacts

Nearly every human activity—from agriculture to transportation to daily living—relies on water resources and affects the availability and quality of those resources. Water resource development has played a role in flood control, agricultural production, industrial and energy development, fish and wildlife resource management, navigation, and a host of other activities. As a result of these impacts, natural hydrologic features have changed through time, pollution has decreased the quality of remaining water resources, and global climate change may affect the distribution of water in the future. Effective and efficient water resource management and planning will need to consider all these factors.

Patterns of water distribution and use. Water availability varies substantially between geographic regions, but it is also affected strongly by the population of the region. Asia, for example, has an extremely large total runoff but the lowest per-capita water availability (**Table 1**). In addition, nearly 40% of the world's population lives in areas that experience severe to moderate water stress. Thus the combination of water and population distribution has resulted in a large difference in per-capita water use between countries (**Fig. 2**).

Total and per-capita water availability by continent

Continent	Total water availability, km^3/year	Per-capita water availability, 1000 m^3/person/year
Europe	2,900	4.23
North and Central America	7,890	17.40
Africa	4,050	5.72
Asia	13,510	3.92
South America	12,030	38.20
Australia and Oceania	2,360	32.20

Worldwide, nearly 4000 km^3 of water is withdrawn every year from surface and ground waters. This is a sixfold increase from the levels withdrawn in 1900 (since which time population has increased four times). Agriculture accounts for the greatest proportion of water use, with about two-thirds of water withdrawals and 85% of water consumption. It also accounts for a great proportion of the increase in water use, with irrigated cropland more than doubling globally since 1960. However, in Europe and

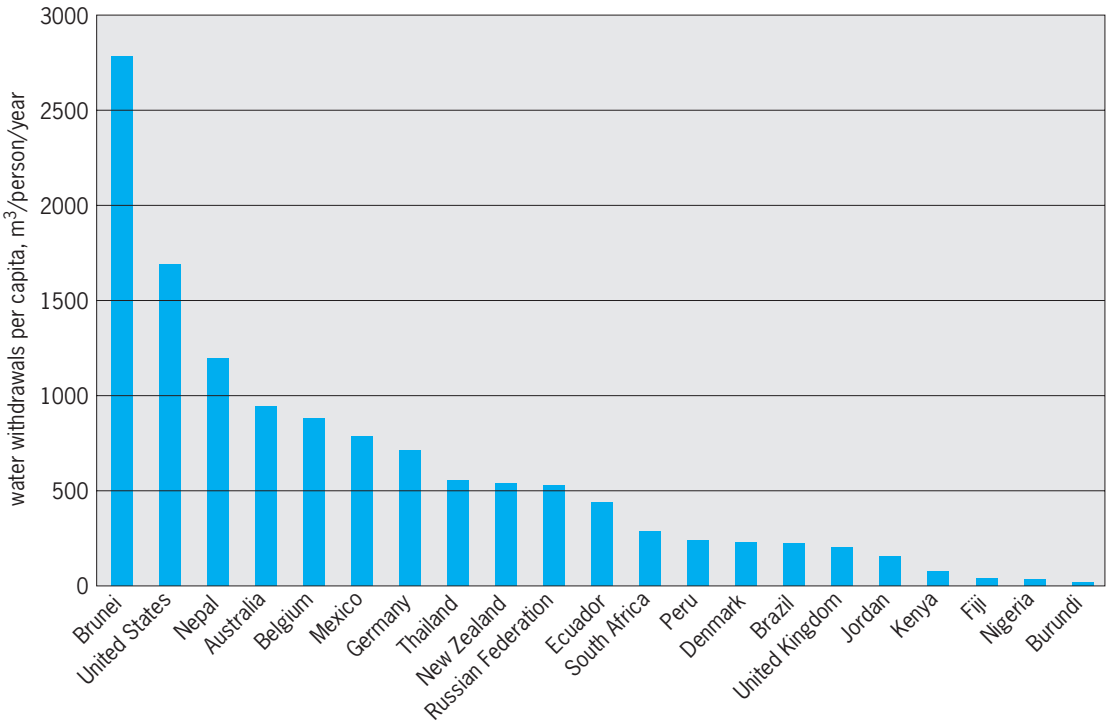


Fig. 2. Estimated per-capita water withdrawals (m^3/person) for selected countries in the year 2000. (*From P. H. Gleick, The World's Water: The Biennial Report on Freshwater Resources, 2000/2001, Pacific Institute for Studies in Development, Environment and Safety, Island Press, Washington, DC, 2001*)

North America particularly, industry consumes a large proportion of available water; industrial uses for water are anticipated to grow on other continents as well.

Changes in hydrologic features. Land development has substantially affected the distribution of water resources. It is estimated that one-half of the natural wetlands in the world have been lost in the last century. In some areas, such as California, wetland loss is estimated to be greater than 90%. The vast majority of wetlands have been lost to diking and draining associated with agricultural development, but urban and industrial changes have reduced wetlands as well. River channels have also been altered to enhance irrigation, navigation, power production, and a variety of other human activities.

Ground-water resources have been depleted in the last century, with many aquifers or artesian sources being depleted more rapidly than they can be recharged. This is called ground-water overdraft. In the United States, ground-water overdraft is a serious problem in the High Plains from Nebraska to Texas and in parts of California and Arizona. In fact, major land subsidence due to ground-water overdraft has occurred in these areas as well as in Japan, Mexico, and Taiwan.

All these changes have resulted in alterations in the natural hydrologic cycle and in the availability of water for human use.

Water pollution. Streams have traditionally served for waste disposal. Towns and cities, industries, and mines provide thousands of pollution sources. Pollution dilution requires large amounts of water. Treatment at the source is safer and less wasteful than flushing untreated or poorly treated wastes downstream. However, sufficient flows must be released to permit the streams to dilute, assimilate, and carry away the treated effluents. *See* WATER POLLUTION.

Global climate change. The availability of fresh water is also likely to be affected by global climate change. There is substantial evidence that global temperatures have risen and will continue to rise. Although the precise effects of this temperature risk on water distribution are challenging to predict, most models of climate change do anticipate increased global precipitation. It is likely that some areas, particularly those at mid to high latitudes, will become wetter, but the increased precipitation will be more seasonal than current patterns. Other areas are likely to receive less precipitation than they do currently. In addition, many models predict increases in the intensity and frequency of severe droughts and floods in at least some regions.

These changes will affect natural stream flow patterns, soil moisture, ground-water recharge, and thus the timing and intensity of human demands for fresh-water supplies. Future conservation and management efforts must consider these changes. *See* GLOBAL CLIMATE CHANGE.

Land management. Land management vitally influences the distribution and character of runoff. Inadequate vegetation or surface organic matter; compaction of farm, ranch, or forest soils by heavy vehicles; frequent crop-harvesting operations; repeated burning; or excessive trampling by livestock, deer, or elk all expose the soil to the destructive energy of rainfall or rapid snowmelt. On such lands little water enters the soil, soil particles are dislodged and quickly washed into watercourses, and gullies may form. *See* LAND-USE PLANNING; SOIL CONSERVATION.

Progress and technologies. There are a variety of measures that can be taken to reduce water consumption; many of them have already been proven effective. In the United States, for example, per-capita water usage dropped 20% from 1980 to 1995. In many cases, improvements to existing systems would contribute to water savings. In the United States, an average of 15% of the water in public supply systems (for cities with populations greater than 10,000) is unaccounted for, and presumably lost. In Jordan, it has been estimated that in the relatively recent past, 30–50% of the domestic water supply is lost due to flaws in the delivery network. In Mexico City, it has been estimated that losses from the water supply system are great enough to support a city the size of Rome.

Improvements can also be achieved by changing industrial and agricultural practices. Agricultural water consumption, the largest single use of water in the world, has an estimated overall water use efficiency of 40%. More effective use of water in agricultural systems can be achieved, for example, with more efficient delivery methods such as drip irrigation. More accurate assessment of soil and plant moisture can allow targeted delivery of water at appropriate times. In industrial settings, recycling and more efficient water use has tremendous potential to reduce water consumption. This is particularly true in the energy sector, where water is used for cooling. Overall, industrial water usage dropped by 30% in California between 1980 and 1990, with some sectors achieving even greater reductions. Japan has achieved a 25% reduction in industrial water use since the 1970s. Additional potential to reduce this usage still exists even in locations such as California where many conservation measures are already in place.

Residential water consumption can also be reduced through conservation measures. For example, prior to 1994, the typical family of four in the United States used 96 gallons (360 liters) of water each day for toilet flushing. New high-efficiency, low-flow toilets can reduce the water required to flush by 70% or more. Additional savings are possible with efficient faucet fixtures and appliances.

Political Concerns

Water conservation in the United States faces a number of institutional as well as technological challenges.

Water rights. In the United States, early rights to water followed the riparian doctrine, which grants the property owner reasonable use of surface waters flowing past his land unimpaired by upstream

landowners. The drier West, however, has favored the appropriation doctrine, which advocates the prior right of the person who first applied the water for beneficial purposes, whether or not his land adjoins the stream. Rights to ground water are generally governed by the same doctrines. Both doctrines are undergoing intensive study.

State laws generally are designed to protect riparian owners against pollution. States administer the regulatory provisions of their pollution-control laws, develop water quality standards and waste-treatment requirements, and supervise construction and maintenance standards of public service water systems. Some states can also regulate ground-water use to prevent serious overdrafts. Artesian wells may have to be capped, permits may be required for drilling new wells, or reasonable use may have to be demonstrated.

Federal responsibilities consist largely of financial support or other stimulation of state and local water management. Federal legislation permits court action on suits involving interstate streams where states fail to take corrective action following persistent failure of a community or industry to comply with minimum waste-treatment requirements. Federal legislation generally requires that benefits of water development projects equal or exceed the costs. It specifies that certain costs be allocated among local beneficiaries but that most of the expense be assumed by the federal government. In 1955, however, the Presidential Advisory Committee on Water Resources Policy recommended that cost sharing be based on benefits received, and that power, industrial, and municipal water-supply beneficiaries pay full cost. These phases of water resource development present difficult and complex questions, because many imponderables enter into the estimates of probable monetary and social benefits from given projects as well as into the cost allocation aspect.

Watershed control. This approach to planning, development, and management rests on the established interdependence of water, land, and people. Conditions on the land are often directly reflected in the behavior of streamflow and in the accumulation of ground water. The integrated approach on smaller watersheds is illustrated by projects under the Watershed Protection and Flood Prevention Act of 1954 (Public Law 566) as amended by P. L. 1018 in 1956. This act originally applied to floods on the smaller tributaries whose watersheds largely are agricultural, but more recently the application has been broadened to include mixed farm and residential areas. Damages from such frequent floods equal half the national total from all floods. Coordination of structures and land-use practices is sought to prevent erosion, promote infiltration, and retard high flows. The Natural Resources Conservation and Forest Services of the Department of Agriculture administer the program. The Natural Resources Conservation Service cooperates with other federal and state agencies and operates primarily through the more than 2000 soil conservation districts.

River basins may be large and complex watersheds. For example, the Tennessee River Basin comprises 40,000 mi^2 (103,600 km^2) in contrast to the 390-mi^2 (1010-km^2) upper limit specified in Public Law 1018. Basin projects may involve systems of multipurpose storage reservoirs, intensive programs of watershed protection, and improvement and management of farm, forest, range, and urban lands. They may call for scientific research, industrial development, health and educational programs, and financial arrangements to stimulate local initiative. The most complete development to date is the Tennessee River Basin, where well-planned cooperative activities have encompassed a wide variety of integrated land and water developments, services, and research.

Water conservation organizations. Organizations for meeting water problems take various forms. Local or intrastate drainage, irrigation, water-supply, or flood-control activities may be handled by special districts, soil conservation districts, or multipurpose state conservancy districts with powers to levy assessments. Interstate compacts have served limited functions on a regional level. To date, Congress has not given serious consideration to proposals for establishing a special federal agency with powers to review and coordinate the recommendations and activities of development services such as the Corps of Engineers, Bureau of Reclamation, Fish and Wildlife Service, Natural Resources Conservation Service, and Forest Service and to resolve conflicts among agencies and citizen groups. However, Congress has passed legislation that has some effects on water use and quality, including the Clean Water Act, the National Environmental Policy Act, the Resource Conservation and Recovery Act, the Safe Drinking Water Act, and the National Energy Policy Act.

In addition, there are a variety of nongovernmental organizations at the local, state, national, and international levels that seek to implement water conservation measures and educate residential, agricultural, and industrial users about available water conservation technologies.

International agreements. Because watersheds often span political boundaries, many efforts to conserve and manage water will require cooperation between states and countries. Many countries currently have international treaties addressing water allocation and utilization. The United States, for example, has treaties with both Mexico and Canada to deal with specific streams such as the Rio Grande and Colorado, or with boundary waters generally, as provided in the treaty with Canada. Similarly, India and Bangladesh have signed a treaty concerning the water in the Ganges River, and Israel and Jordan have an agreement governing the allocation of water from the Jordan River. In 1997, the United Nations adopted the Convention on the Law of the Non-navigational Uses of International Watercourses, which includes an obligation not to cause significant harm to other watercourse states, as well as provisions for dispute resolution. In addition, in 1996 the Global Water Partnership and the

World Water Council were formed for the purpose of addressing ongoing international water concerns. Bernard Frank; Michelle McClure; George Press

Coastal Water Conservation

Most coastal waters less than 300 ft (100 m) deep were dry land 15,000 years ago. The North Sea was a peat bog, for example, and one could walk from England to France or from Siberia to Alaska. As the glaciers retreated, these exposed continental shelves began to fill with water until they now constitute 10% of the world's ocean area. The average depth of the present continental shelf is 183 ft (60 m), with a width extending 47 mi (75 km) from shore. The salinity of coastal water ranges from 35% (100% ocean water) at the seaward edge of the continental shelf (660 ft or 200 m depth contour) to 0% (100% fresh water) within coastal estuaries and bays at the shoreward edge of the shelf. The annual temperature range of mid-latitude shelf water is 68°F (20°C) off New York and 50°F (10°C) off Oregon, with less temperature change in tropical waters. *See* CONTINENTAL MARGIN.

Resources. Because of the shallow bottom, compared with the deep ocean of 16,500 ft (5000 m) depth, organic matter is transformed to nutrients and recycled (returned to the water column) faster on the continental shelves. The growth of plants in the sunlit regions of these relatively shallow areas is thus 10 times that of the open ocean, and the rest of the shelf food web is similarly more productive. Approximately 99% of the world's fish catch is taken from these rich shelves. As a result of their accessibility and commercial value, coastal waters have been the object of extensive scientific studies.

Withdrawal of the Wisconsin glacier and buildup of the native Amerindic populations about 10,000 years ago led to simple harvesting of the living resources of the United States continental shelves. Since colonial days, however, this coastal region has been the focus of increased exploitation with little thought given to the impact of these activities. After the discovery of codfish in the New World by Cabot in 1497, a "foreign" fishing fleet was inaugurated by the French in 1502, the Portuguese in 1506, the Spanish in 1540, and the English in 1578; the first "domestic" fishery of the United States was initiated by the ill-fated Roanoke colony in 1586. The adjacent human population then grew from a few Indian settlements scattered along the coast to the present east coast megalopolis, housed in an almost continuous urban development from Norfolk, Virginia, to Portland, Maine. By 1970, continued fishing pressure of the foreign and domestic fleets had reduced the fish stocks of the northeast continental shelf to approximately 25% of their virgin biomass. *See* MARINE FISHERIES.

Pollutant impacts. At the same time, attempts at waste control in colonial days began as early as 1675 with a proclamation by the governor of New York against dispersal of refuse within the harbor, yet the New York urban effluent expanded until the percent saturation of dissolved oxygen of the harbor halved between 1910 and 1930. The amount of trace metals in the New York Bight Apex sediments now exceeds that of the outer shelf by as much as a hundredfold. Questions about the impact of extended offshore United States jurisdiction of fisheries, construction of ocean sewage outfalls, dredging, beach erosion, and emplacement of pipelines are hotly debated issues in coastal communities that depend on revenue from commercial fishing, tourism, and other forms of recreational activities. As a result of these possibly conflicting uses of the coastal zone, multidisciplinary research on this ecosystem has been intensified over the years by the U.S. Department of Energy (DOE), the National Oceanic and Atmospheric Administration (NOAA), the U.S. Geological Survey (USGS), the Environmental Protection Agency (EPA), the Bureau of Land Management (BLM), and the National Science Foundation (NSF).

People have come to realize that dilution of wastes by marine waters can no longer be considered a simple or permanent removal process within either the open ocean or nearshore waters. The increasing utilization of the continental shelf for oil drilling and transport, siting of nuclear power plants, and various types of planned and inadvertent waste disposal, as well as for food and recreation, requires careful management of human activities in this ecosystem. Nearshore waters are presently subject, of course, to both atmospheric and coastal input of pollutants in the form of heavy metals, synthetic chemicals, petroleum hydrocarbons, radionuclides, and other urban wastes.

However, overfishing is an additional human-induced stress. For example, the sardine fishery collapsed off the California coast, herring stocks are down off the east coast, and the world's largest fishery, for anchovy off Peru, has been reduced to less than 10% of its peak harvest in the late 1960s. Determination of what is the cause and which is the direct effect within a perturbation response of the food web of this highly variable continental shelf ecosystem is a difficult matter. One must be able to specify the consequences of human impact within the context of natural variability; for example, populations of sardines exhibited large fluctuations within the geological record off California before a fishery was initiated.

Furthermore, physical transport of pollutants, their modification by the coastal food web, and demonstration of transfer to humans are sequential problems of increasing complexity on the continental shelf. For example, after 30 years of discharge of mercury into the sea, the origin of the Minimata neurological disease of Japan was finally traced to human consumption of fish and shellfish containing methyl mercuric chloride. The Itai itai disease is now attributed to ingestion of food with high cadmium levels. Discharges of chlorinated hydrocarbons, such as DDT off California, polychlorinated biphenyl (PCB) in both the Hudson River, New York, and within Escambia Bay, Florida, mirex in the Gulf of Mexico, and vinyl chloride in the North Sea, have also led to inhibition of algal photosynthesis, large mortality of

shrimp, and reproductive failure of birds and fish.

Oil spills constitute an estimated annual input of 2×10^6 tons of petroleum to the continental shelves with an unresolved ecological impact; another 2×10^6 tons of petrochemicals is added each year from river and sewer runoff. Fission and neutron-activation products of coastal reactors, such as San Onofre (California), Hanford (Washington), and Wind-scale (United Kingdom), are concentrated in marine food chains with, for example, cesium-137 found in muscle tissue of fish, ruthenium-106 in seaweed, zinc-65 in oysters, and cerium-144 in phytoplankton their somatic and genetic effects on humans are presumably minimal. Finally, disposal of dissolved and floatable waste material from New York City has been implicated as a possible factor in both shellfish loss off New Jersey and the occasional closure of Long Island beaches.

Simulation models. One approach to quantitatively assess the above pollutant impacts is to construct simulation models of the coastal food web in a systems analysis of the continental shelf. Models of physical transport of pollutants have been the most successful, for example, as in studies of beach fouling by oil. Incorporation of additional biological and chemical terms in a simulation model, however, requires dosage response functions of the natural organisms to each class of pollutants, as well as a quantitative description of the "normal" food web interactions of the continental shelf.

Toxicity levels in terms of median lethal concentrations (LC50) of metals, pesticides, biofouling agents (such as chlorine), PCB, and petroleum fractions have been determined only for organisms that can be cultured in the laboratory. The actual form of the pollutant, such as methyl mercuric chloride or chloramine, and its concentration in the marine environment, however, are not always known. Furthermore, the actual contribution of a pollutant to mortality of organisms within a coastal food web is additionally confounded by the lack of understanding of natural mortality. Natural death on the continental shelf is a poorly known process. Nevertheless, there are some clear-cut examples of pollutant impacts on the coastal zone and, in these cases, management decisions are continually being made to correct these situations. *See* FOOD WEB.

Sewage. For example, raw sewage contains pathogenic bacteria that cause human diseases such as typhoid, typhus, and hepatitis. These and various gastroenteric diseases may be contracted from eating raw shellfish that live in sewage-polluted waters. Health authorities have closed more than a million acres (4000 km^2) of the best shellfish beds and put segments of the shellfish industry abruptly out of business, with economic losses running to tens of millions of dollars per year. Purification of sewage is thus absolutely necessary for healthy coastal waters. The cost increases as the degree of treatment is intensified, however. Most treatments remove only a fraction of dissolved fertilizing minerals such as nitrates and phosphates. These nutrients from sewage plants overfertilize coastal waters where the effluent is discharged, and can at times lead to oxygen depletion of bottom waters. The cost of this additional removal of nutrients must now be weighed against their potential damage to the coastal ecosystem. *See* SEWAGE TREATMENT.

Toxic materials. Insecticides also reach coastal waters via runoff from the land, often causing fish kills. Any amount above one-tenth part of insecticide to a million parts of water can be lethal to some fish for most of the following: DDT, parathion, malathion, endrin, dieldrin, toxaphene, lindane, and heptachlor. Contamination of fish eggs by DDT is fatal to a high proportion of young. Insecticides function mainly as paralytic nerve poisons with resulting lack of coordination, erratic behavior, loss of equilibrium, muscle spasms, convulsions, and finally suffocation. Federal and state legislation has all but eliminated DDT from future use in the United States. *See* INSECTICIDE; PESTICIDE.

Other chemical pollutants such as metals, acids, and gases result from industrial activities. Paper and pulp mills discharge wastes that are dangerous to aquatic life because the wastes have a high oxygen demand and deplete oxygen. Other factories discharge lead, copper, zinc, nickel, mercury, cadmium, and cobalt, which are toxic to coastal life in concentrations as low as 0.5 part per million (ppm). Cyanide, sulfide, ammonia, chlorine, fluorine, and their combined compounds are also poisonous. To prevent chemical pollution of the environment, factories are required to remove contaminants from their wastes before discharging them into coastal waters or into local sewage systems.

Oil pollution arises from various sources. Most cases of fish poisoning are from accidental spillage from tankers, storage depots, or wells. However, slow but constant leakage from refineries ruins waterways and is difficult to remedy. Oysters seem unable to breed in the vicinity of refineries. Enclosed ocean regions take longer to recover from oil spills than open coastal areas. Careless handling at plants also results in water pollution by poisonous byproducts, such as cresols and phenols that are toxic in amounts of 5–10 ppm. In past years tankers used to pump oil into the water while cleaning their tanks, but this and cleanup procedures after oil spills are being corrected by stronger federal laws.

Thermal pollution. Thermal pollution is caused by the discharge of hot water from power plants or factories and from desalination plants. Power plants are the main source of heated discharges. They are placed at the coast or on bays to secure a ready source of seawater coolant. A large power installation may pump in 10^6 gal/min (63 m^3/s) and discharge it at a temperature approximately 18°F (10°C) above that of the ambient water. Although temperatures of coastal waters range from summer highs of 95°F (35°C) in southern lagoons to winter lows of 30.3°F (−1°C) in northern estuaries, each has a typical pattern of seasonal temperature to which life there has adapted. In a shallow bay with restricted tidal flow, the rise in temperature can cause gross alterations to the natural ecology. Federal standards prohibit heating of

coastal waters by more than 0.9°F (0.5°C). *See* THERMAL ECOLOGY.

Dredging. Finally, dredging waters to fill wetlands for house lots, parking lots, or industrial sites destroys the marshes that provide sanctuary for waterfowl and for the young of estuarine fishes. As the bay bottom is torn up, the loosened sediments shift about with the current and settle in thick masses on the bottom, suffocating animals and plants. In this way, the marshes are eliminated and the adjoining bays are degraded as aquatic life zones. The northeast Atlantic states have lost 45,000 acres (182 km^2) of coastal wetlands in only 10 years, and San Francisco Bay has been nearly half obliterated by filling. Dredging to remove sand and gravel has the same disruptive effects as dredging for landfill or other purposes, whether the sand and gravel are sold for profit or used to replenish beach sand eroded away by storms. The dredging of boat channels adds to the siltation problem, and disposal of dredge spoils is being regulated in coastal areas.

Management for the future. Human populations have grown to a level where they now can have serious impacts on coastal waters. Past experience suggests that human-induced stress is most likely to lead to species replacement by undesirable forms rather than a decrease in the organic production of the ecosystem. Any societal action must now be considered in the context of what is known about the shelf ecosystem, what management decision is required, what perturbation events are likely to ensue, and what the societal costs are in using renewable and nonrenewable coastal water resources. Prediction of such perturbation events has both immediate and future value to humans in terms of management and conservation options, such as removal of shellfish before depletion of bottom oxygen, the best mode of sewage treatment, preservation of coastal species, and a decrease of toxicant levels within the coastal food web. As one moves from prediction of meteorological events to biological changes of the coastal food web, however, increasing sources of error emerge in the predictions. Since the mid-1800s, humans have introduced more sources of environmental variability to the continental shelf than this coastal ecosystem has encountered during the last 10,000 years. Nevertheless, sufficient information on continental shelf processes is emerging to suggest that specification of management options by delineation of cause and effect within a perturbation response of the coastal zone is a feasible goal.

John J. Walsh

Bibliography. P. E. Black, *Conservation of Water and Related Land Resources*, 2d ed., 1988; T. Dunne and L. B. Leopold, *Water in Environmental Planning*, W. H. Freeman, New York, 1978; P. H. Gleick, *The World's Water: The Biennial Report on Freshwater Resources, 2000/2001*, Publication of the Pacific Institute for Studies in Development, Environment and Safety, Island Press, Washington, DC, 2001; A. W. Hornslow, *Water Quality Data: Analysis and Interpretation*, 1995; N. T. Kottegoda, *Stochastic Water Resources Technology*, 1980; J. W. Moore, *Balancing the Needs of Water Use*, 1988; Soil and Water Conservation Society, *Sustainable Agricultural Systems*, 1990; F. Van der Leeden, F. L. Troise, and D. K. Todd, *The Water Encyclopedia*, 2d ed., Geraghty and Miller Groundwater Series, Lewis Publishers, Chelsea, MI, 1990.

Water desalination

The removal of dissolved minerals (including salts) from seawater or brackish water. This may occur naturally as part of the hydrologic cycle or as an engineered process. Engineered water desalination processes, producing potable water from seawater or brackish water, have become important because many regions throughout the world suffer from water shortages caused by the uneven distribution of the natural water supply and by human use. The capacity of installed desalination plants around the world at the end of 1966 was 200,000 m^3/day (53,000,000 gal/day); by 2002, it had increased to 32,400,000 m^3/day (8,560,000,000 gal/day). *See* WATER SUPPLY ENGINEERING.

Seawater, brackish water, and fresh water have different levels of salinity, which is often expressed by the total dissolved solids (TDS) concentration. Seawater has a TDS concentration of about 35,000 mg/L, and brackish water has a TDS concentration of 1000–10,000 mg/L. Water is considered fresh when its TDS concentration is below 500 mg/L, which is the secondary (voluntary) drinking water standard for the United States. Salinity is also expressed by the water's chloride concentration, which is about half of its TDS concentration. *See* SEAWATER.

Water desalination processes separate feedwater into two streams: a fresh-water stream with a TDS concentration much less than that of the feedwater, and a brine stream with a TDS concentration higher than that of the feedwater.

Thermal processes. Distillation is a process that turns seawater into vapor by boiling, and then condenses the vapor to produce fresh water. Boiling water is an energy-intensive operation, requiring about 4.2 kilojoules of energy (or latent heat) to raise the temperature of 1 kg of water by 1°C. After water reaches its boiling point, another 2257 kJ of energy (or the heat of vaporization) is required to convert it to vapor. The boiling point depends on ambient atmospheric pressure—at lower pressure, the boiling point of water is lower. Therefore, keeping water boiling can be accomplished either by providing a constant energy supply or by reducing the ambient atmospheric pressure. *See* DISTILLATION.

Multistage flash distillation (MFD). The MFD process, developed in the 1960s, is still one of the most popular desalination processes. An MFD system consists of serially arranged cylindrical tanks, called flash chambers, in which the pressure is reduced (**Fig. 1**). Seawater is discharged into the first chamber and pressure reduction lowers the water's boiling point, causing vapor to "flash off" the water surface. The water vapor is then condensed, and the resulting

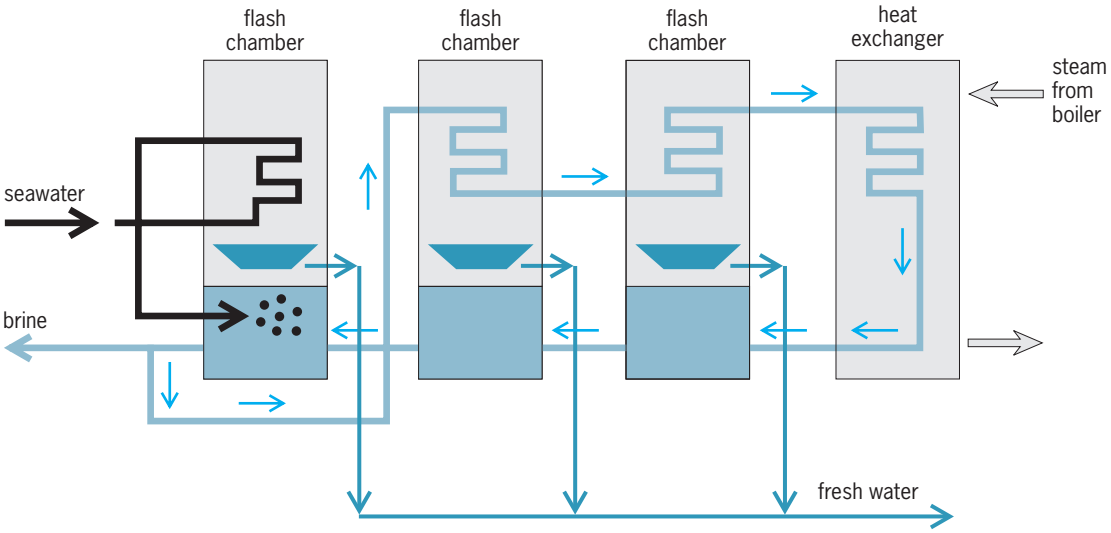


Fig. 1. Multistage flash distillation process.

fresh water is collected. The outflow of concentrated brine from one chamber becomes the inflow of the next chamber, and so on. In this process, a corresponding lower pressure must be maintained in each subsequent chamber for flashing-off to occur.

Multieffect distillation (MED). MED, an alternative distillation process, is used in places where high-temperature steam is available at reasonable cost. In an MED unit, feedwater flows into the top of a series of evaporators and falls as a thin film down the inside of long vertical tubes. The steam surrounding these tubes causes the feedwater to boil as it falls. The resulting vapor is led away, while brine collects at the bottom of the unit. Part of the feed steam condenses to a fresh-water product. The surface of long vertical tubes of all the other units is heated by the steam produced in each preceding unit. *See* EVAPORATION; EVAPORATOR; STEAM.

One major problem with both the MFD and MED processes is scale formation. Substances such as calcium sulfate in feedwater have low solubility in warmer water, and precipitate as the temperature rises, forming scales on the equipment surface. A lower operating temperature can reduce the scale problem, but not without decreasing the thermal efficiency. More research and development on scale control is needed.

Vacuum freezing. When seawater is frozen by reducing the pressure and temperature, the dissolved minerals or salts separate from the ice crystals. After freezing, the ice crystals are washed to remove the brine and melted to produce fresh water. In principle, freezing desalination processes should consume less energy than distillation processes, but no freezing process has been commercially successful. In terms of process engineering, it is more difficult to handle ice than steam.

Membrane processes. When a salt solution is separated from pure water by a semipermeable membrane, water tends to diffuse through the membrane into the salt solution. This is the well-known natural phenomenon called osmosis. *See* OSMOSIS.

Reverse osmosis. Reverse osmosis, the process that causes water in a salt solution to move through a semipermeable membrane to the freshwater side, is accomplished by applying pressure in excess of the natural osmotic pressure to the salt solution (**Fig. 2**). The operational pressure of reverse osmosis for seawater desalination is much higher than that

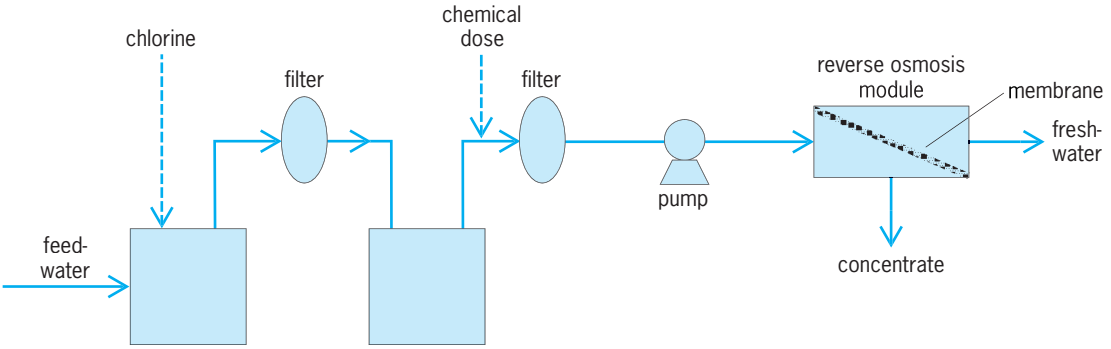


Fig. 2. Flow chart of reverse osmosis desalination.

for brackish water, as the osmotic pressure of seawater at a TDS concentration of 35,000 mg/L is about 2700 kPa (400 psi) while the osmotic pressure of brackish water at a TDS concentration of 3000 mg/L is only about 230 kPa (30 psi).

Electrodialysis. Salts dissociate into positively and negatively charged ions in water. The electrodialysis process uses semipermeable and ion-specific membranes, which allow the passage of either positively or negatively charged ions while blocking the passage of the oppositely charged ions. An electrodialysis membrane unit consists of a number of cell pairs bound together with electrodes on the outside. These cells contain an anion exchange membrane and cation exchange membrane (**Fig. 3**). Feedwater passes simultaneously in parallel paths through all of the cells, separating the product (water) and ion concentrate. *See* DIALYSIS; ION EXCHANGE.

Electrodialysis was successfully introduced in the early 1960s (about 10 years before the introduction of reverse osmosis) as a commercial means of desalting brackish water. This method is cost effective because the unit can be operated at low temperature and pressure.

Membrane fouling. A major problem in the membrane processes of water desalination is fouling, which is the scaling or plugging of membrane surfaces over time by organic and inorganic substances in the feedwater. Fouling prevention requires the pretreatment of the feedwater or the addition of antiscalants. There are three membrane-cleaning methods. In hydraulic cleaning, the flow direction is changed (backflushing) to remove fouling at the membrane surface. In mechanical cleaning, the membrane is cleared with sponge balls. In chemical cleaning, the fouled membrane is washed with chemical agents such as acid for mineral scale or alkali for organic matter. New fouling-resistant membrane materials are being developed by studying the physicochemical and biological interactions between membrane surface and foulants and antifouling agents. *See* MEMBRANE SEPARATIONS.

Processes using natural energy. Desalination processes are energy-intensive. Research continues for improving the energy efficiency of existing processes (which use conventional energy), and new technology is being developed for water desalination processes that are driven by natural energy.

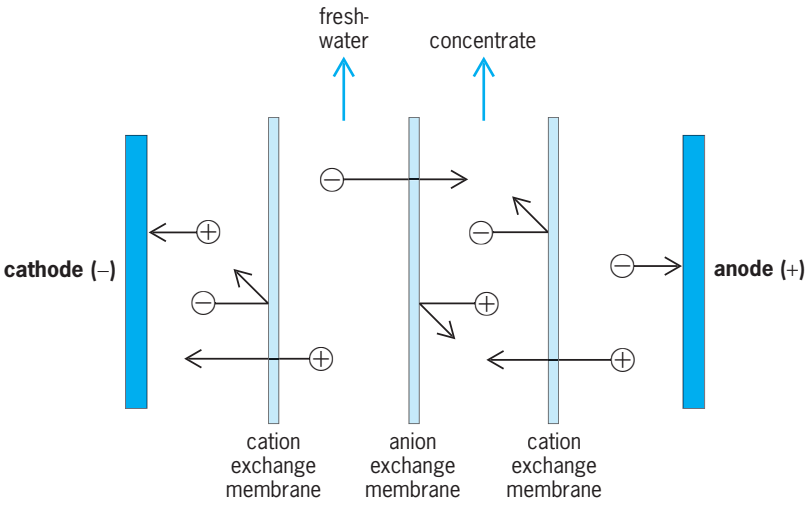


Fig. 3. Electrodialysis process.

Open-cycle OTEC system. Ocean thermal energy conversion (OTEC) uses the ocean's natural thermal gradient between surface water and water at a depth of 1000 m (3300 ft) or more to drive a power-producing cycle. In an open-cycle OTEC system, warm seawater is the working fluid. The warm seawater is flash-evaporated in a vacuum chamber to produce steam. The steam expands through a turbine that is coupled to a generator to produce electricity. The steam existing the turbine is condensed to fresh water, using cold, deep-ocean water (**Fig. 4**).

Wind-powered brackish-water desalination. At present, most interest in wind energy is focused on converting it to electric power. However, converting wind power to electric power and then converting the electric power to hydraulic power for water desalination is not energy efficient or cost effective. A prototype wind-powered reverse-osmosis system, which converts wind power directly to hydraulic power of feedwater into the reverse-osmosis module, was developed by researchers at the University of Hawaii. This prototype system was successfully tested on

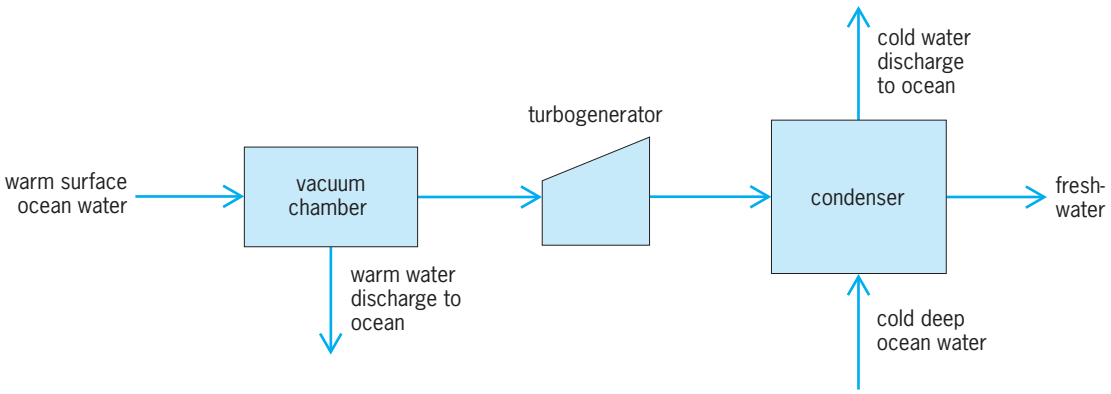


Fig. 4. Open-cycle OTEC system.

Coconut Island off the windward coast of Oahu, Hawaii. *See* WIND POWER.

Wave-driven seawater desalination. An ocean wave–powered reverse-osmosis desalination system was developed at the University of Delaware. This system, which consists of a wave pump and a reverse-osmosis module, was successfully tested at two Caribbean locations. *See* ENERGY SOURCES.

Clark C. K. Liu; Jae-Woo Park

Bibliography. D. Block and J. Valenzuela, *Thermoeconomic Optimization of OC-OTEC Electricity and Water Production Plants*, Solar Energy Research Institute, Golden, CO, SERI/STR-251-2603, 1985; R. G. Gutman, *Membrane Filtration*, Adam Hilger, Bristol, England, 1987; D. Hicks et al., DELBUOY: Ocean wave-powered seawater reverse osmosis desalination system, *Desalination*, 73:81–94, 1989; M. Mulder, *Basic Principles of Membrane Technology*, 2d ed., Kluwer Academic, Dordrecht, 1996; National Research Council, *Review of the Desalination and Water Purification Technology Roadmap*, National Academies Press, 2004; K. Wangnick, *1998 IDA Worldwide Desalination Plants Inventory*, Rep. 15, International Desalination Association, Topsfield, MA, 1998.

Water hammer

The propagation in a liquid of an acoustic wave that is caused by a rapid change in fluid velocity. Such relatively sudden changes in the liquid velocity are due to events such as the operation of pumps or valves in pipelines, the collapse of vapor bubbles within the liquid, underwater explosions, or the impact of water following the rapid expulsion of air from a vent or a partially open valve. Alternative terms such as pressure transients, pressure surge, hydraulic transients, and hydraulic shock are often employed. Although the physics and mathematical characterization of water hammer and underwater acoustics (employed in sonar) are identical, underwater sound is always associated with very small pressure changes compared to the potential of moderate to very large pressure differences associated with water hammer. *See* CAVITATION; SOUND; UNDERWATER SOUND.

A pressure change Δp is always associated with the rapid velocity change ΔV across a water hammer wave, as formulated from the basic physics of mass and momentum conservation by the Joukowsky equation, $\Delta p = -\rho a \Delta V$. Here ρ is the liquid mass density and a is the sonic velocity of the pressure wave in the fluid medium. In a pipe, this velocity depends on the ratio of the bulk modulus of the liquid to the elastic modulus of the pipe wall, and on the ratio of the inside diameter of the pipe to the wall thickness. In water in a very rigid pipe or in a tank, or even the sea, the acoustic velocity is approximately 1440 m/s (4720 ft/s), a value many times that of any liquid velocity.

For an extremely sudden change of the flow of water (assumed instantaneous) in a steel pipe, the effect of water hammer is dramatic (see **table**). However, the acoustic velocity and the pressure rise associated with water hammer are somewhat reduced by the elasticity of the pipe wall. For very elastic pipes such as polyethylene, the acoustic velocity, and the resultant pressure rise for each unit of velocity extinction, is reduced by a much greater amount. *See* ELASTICITY.

Acoustic velocity and pressure change for water in steel pipe*

Inside diameter of pipe ÷ wall thickness	Acoustic velocity (a), m/s (ft/s)	Pressure change (Δp), kPa (psi)
20	1310 (4300)	399 (57.8)
40	1210 (3970)	368 (53.4)
60	1131 (3710)	344 (49.9)
80	1064 (3490)	324 (46.9)

*For velocity change (ΔV) of 0.3048 m/s (1 ft/s).

Fortunately, liquid-handling systems in engineering are designed so that water hammer does not result from sudden closure, but is limited to more gradual flow changes initiated by valves or other devices. The dramatic pressure rise (or drop) results (see table) can be significantly reduced by reflections of the original wave from pipe-area changes, tanks, reservoirs, and so forth. Although the Joukowsky equation applies across every wavelet, the effect of complete valve closure over a period of time greater than a minimum critical time can be quite beneficial. This critical time is the time required for an acoustic wave to propagate twice the distance along the pipe from the point of wave creation to the location of the first pipe-area change. *See* HYDRODYNAMICS; PIPE FLOW.

C. Samuel Martin

Bibliography. M. H. Chaudhry, *Applied Hydraulic Transients*, 2d ed., 1987; F. J. Moody, *Introduction to Unsteady Thermofluid Mechanics*, 1990; B. B. Sharp, *Water Hammer: Practical Solutions*, 1996; E. B. Wylie and V. L. Streeter, *Fluid Transients in Systems*, 1993.

Water-jet cutting

The use of high-pressure water jets, which may contain abrasive powder, for cutting and removing materials. For example, water accelerated up to twice the speed of sound [343 m/s (1125 ft/s) at 20°C (68°F)] can penetrate and cut rock in a few seconds.

Among the methods of cutting metal and nonmetallic materials, pure and abrasive water-jet cutting techniques have a distinct advantage because of their versatility and speed (see **table**). They can cut all materials, including hard-to-machine materials such as super-alloy, Kevlar, and boron carbide. They can also easily cut aerospace materials such as graphite composite and titanium, and brittle materials such as advanced ceramics, granite, marble, and glass (**Fig. 1**). The pure water-jet technique with a relatively high flow rate (2–10 gal/min or 8–38 L/min) is

Cutting speeds of abrasive water jets

Material	Thickness in.	Thickness mm	Speed in./min	Speed mm/min
Aluminium	0.25	6.25	9–40	225–1000
	1.00	25	1–10	25–250
Steel	0.25	6.25	5–20	125–500
	1.00	25	0.5–5	12.5–750
Titanium	0.25	6.25	5–30	125–750
	1.00	25	1–7	25–175
Superalloy	0.25	6.25	8–15	200–375
	1.00	25	0.1–0.5	2.5–12.5
Kevlar	1.00	25	2–10	50–250
Graphite composite	1.00	25	3–10	75–250

widely used by the construction industry for applications such as road repair and tunnel boring. It is also used by the food industry to cut candy and chocolate bars, meats, vegetables, and fruits. Precision enhancement of the technology has led to medical applications, including orthopedic surgery for hip joint replacement. Other biomedical applications include removing clots from blood vessels and a micro water jet for corneal surgery.

The advantages of pure and abrasive water-jet cutting are (1) absence of thermal distortion and work hardening; (2) noncontact during cutting, thus eliminating tool wear and contact force; and (3) omnidirectional cutting, allowing the cutting of complex shapes and contours.

Although the use of the water-jet system is rapidly growing, the technique has some drawbacks and limitations. Water-jet technology has not yet developed fully for high-tolerance and -precision machining. The initial capital investment for the system, including the motion-control equipment and operating costs, is relatively high. And the noise level (80 adjusted decibels) is somewhat high, but the system can be specially designed to isolate the noise source. *See* MACHINABILITY OF METALS.

Cutting principle. The water-jet pump and its delivery system are designed to produce a high-velocity jet stream within a relatively short trajectory distance, since the kinetic energy of the water and abrasive particles is directly proportional to the square of the jet velocity. In abrasive jet cutting applications, the abrasives entrained in the jet stream usually attain approximately 80% of the water-droplet velocity at the nozzle tip. The jet cuts the material by a rapid erosion process, when its force exceeds the compressive strength of the material. Erosion mechanics is a highly dynamic and material-dependent phenomenon involving shear and tensile failure due to localized stress fields. Since the area eroded by the abrasive is also swept by the water stream, the heat generated during the cutting is dissipated immediately, resulting in a small rise in temperature (less than 90°F or 50°C) in the workpiece. Therefore, no thermal distortion or work hardening is associated with water-jet cutting. The cutting by rapid erosion also significantly reduces the actual force exerted on the material, enabling the water jet to cut fragile or deformable materials such as glass and honeycomb structures. Unlike traditional cutting and machining such as turning and drilling, where the cutter is fed by a continuous and constant level of energy during the entire cutting operation, the abrasive water-jet stream loses energy along its cutting path. The cutting power of the jet stream decreases from the top of the material to the bottom, leaving a tapered kerf (slit) and striation marks on the lower portion of the cut surface. This phenomenon is typical of high-energy beam-cutting applications such as laser and electron-beam cutting. *See* JET FLOW; METAL, MECHANICAL PROPERTIES OF; SHEAR.

Operation. A typical ultrahigh-pressure water-jet cutting system consists of two stages of pumping equipment, a high-pressure water-jet delivery system, a mechanical manipulator for motion control, and a discharge catcher system.

Pump. The first stage of pumping is carried out by a radial displacement pump, which can pressurize the hydraulic oil up to 3000 lb/in.^{2} (20.7 megapascals). The hydraulic oil then drives an intensifier pump (**Fig. 2**). The typical intensification ratio of this piston type of pump is 1:20, which is the area ratio of two ends of the piston. The intensifier pump can pressurize water up to 60,000 lb/in.^{2} (414 MPa). The tap water supplied to the intensifier pump passes through several stages of filters (10-0.5-micrometer range) at a boosted water-feed pressure of 80 lb/in.^{2} (0.555 MPa). The high-pressure water from the intensifier pump is channeled through an accumulator to level off pressure fluctuations created by the plunger motion of the pump. The pressurized water is then carried to a cutting station by means of stainless-steel high-pressure tubing and swivel joints or by a flexible high-pressure hose which can withstand pressures

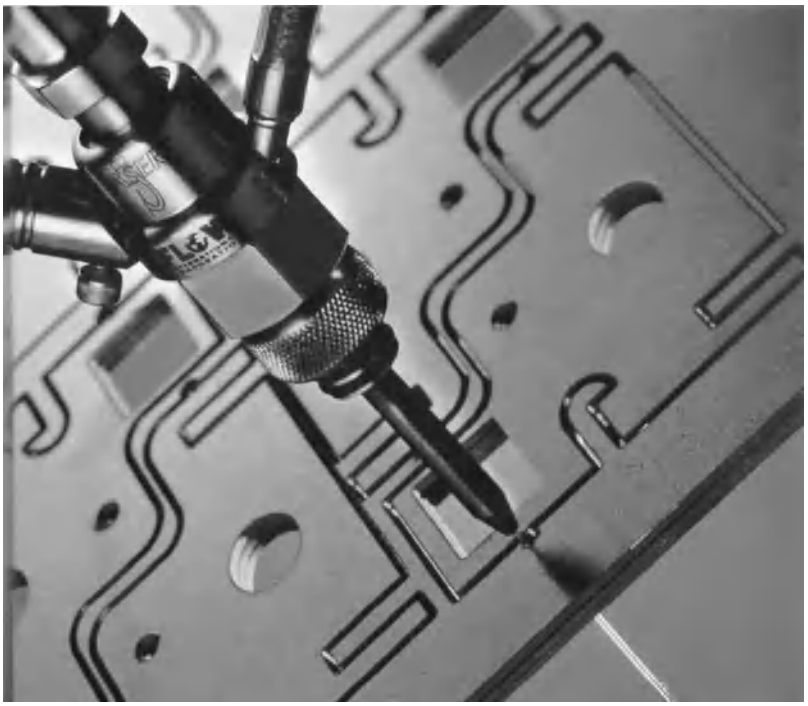


Fig. 1. Abrasive water-jet cutting of 0.5-in.-thick (12.5-mm) titanium at a pressure of 45,000 lb/in.^{2} (310 MPa). (*Flow International Corp.*)

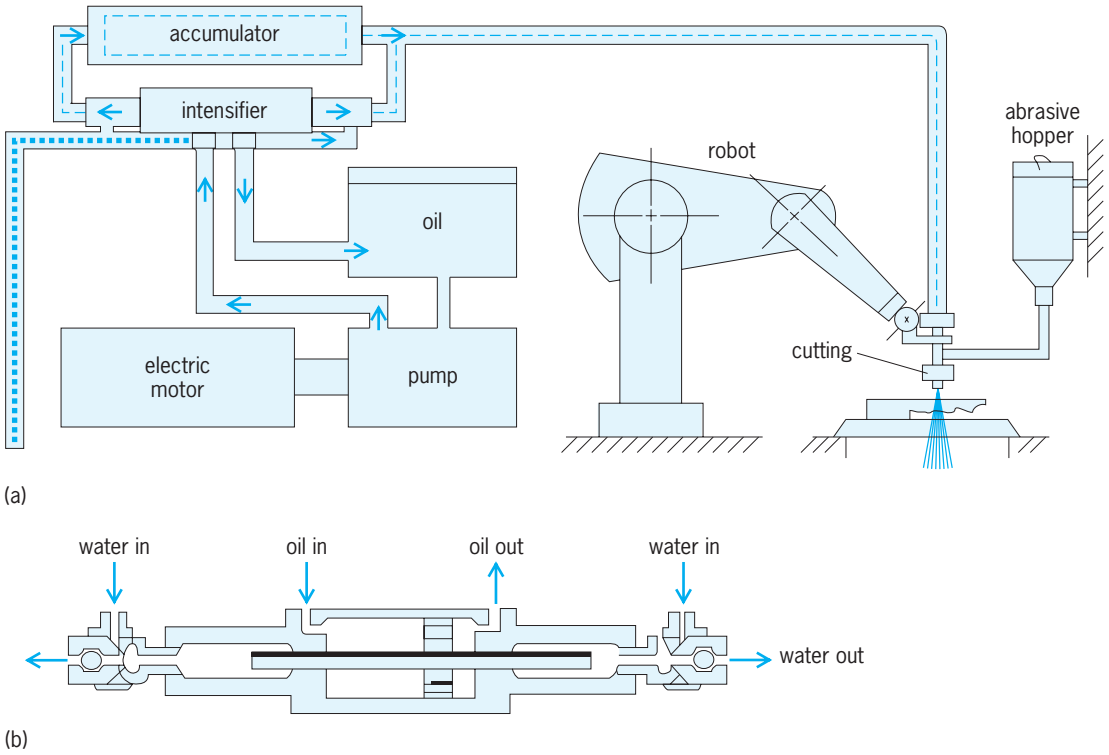


Fig. 2. Water-jet cutting system. (*a*) Water-jet pump and motion control system. (*b*) Double-acting intensifier pump.

up to 100,000 lb/in.^{2} (690 MPa). *See* DISPLACEMENT PUMP.

Jet delivery system. The high-pressure water delivered to a nozzle assembly is first converted into a high-speed jet through an orifice assembly which houses an on–off control valve and orifice. The jet created through the orifice can be manipulated by motion-control equipment such as a robot or *x-y* table to cut a desired shape of workpiece. For cutting hard and dense materials such as steel, stone, and ceramics, the water jet is mixed with abrasive particles in a mixing chamber (**Fig. 3**) before being discharged through the nozzle. The orifice is made of sapphire or diamond, and it lasts up to 200 and 2000 h, respectively. The nozzle (often called the mixing tube) is made of tungsten carbide, which has a life of 5 h under normal operating conditions. A new nozzle material, a composite carbide, can extend the life of the nozzle up to 65 h. The coherent water jet emerging from the tip of the nozzle can attain a speed of 2000 ft/s (610 m/s) at a normal operating pressure of 45,000 lb/in.^{2} (310 MPa). Recent advances in high-pressure pump technology have led to the development of several new single-stage pumps, which also can deliver up to 45,000 lb/in.^{2} but with higher flow rates.

Process control parameters. The precision and efficiency of water-jet cutting can be controlled by the process parameters, including the water pressure; abrasive type, size, and flow rate; orifice and nozzle size; standoff distance; cutting angle; traverse rate; and target material strength (Fig. 3). The velocity of the jet stream as well as the velocity of the entrained abrasive (usually 20% less than the water-droplet velocity) is primarily governed by the water pressure. The jet velocity v can be calculated from the equation below, where P is the water pressure, ρ is the water

$$V = K\sqrt{\frac{2P}{\rho}}$$

density, and K is a dimensionless system constant, of order unity, which includes the effects of water compressibility, orifice efficiency, and the abrasive and

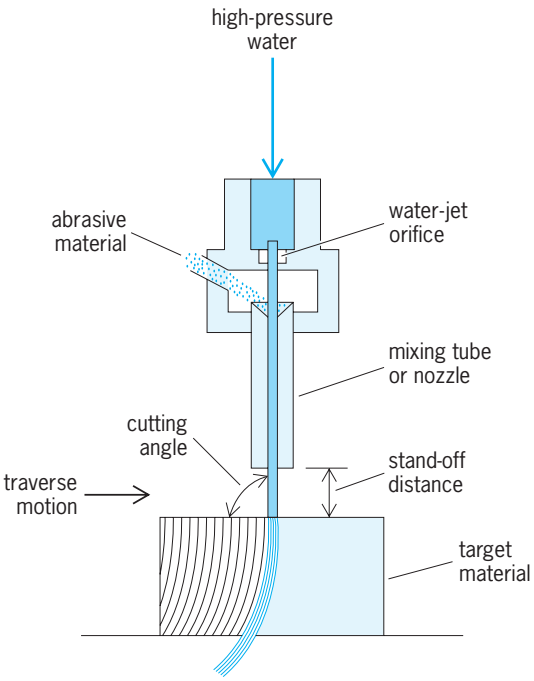


Fig. 3. Typical abrasive water-jet cutting head.

water mass flow rates. The water flow rate is controlled by the water pressure and orifice configuration. The cutting efficiency generally increases as the water flow rate increases. Since the water flow rate is limited by the pump capacity, the orifice size for typical commercially available pumps (20–100 hp, 15–75 kW) ranges between 0.004 and 0.022 in. (0.1 and 0.55 mm). The corresponding range of water flow rate is 0.5–2 gal/min (1.89–7.57 L/min). The optimized pairing in size selection for orifice and nozzle is also an important factor in increasing the efficiency of the cutting operation. The optimum pairing ratio (orifice diameter/nozzle diameter) is 0.3–0.4.

Abrasives. A mixture of water and abrasives is used for cutting hard materials. Garnet is the most commonly used abrasive in industrial cutting operations. Other abrasives such as aluminum oxide, silicon carbide, silica sand, glass bead, and steel grit are also used in special-purpose cutting and cleaning operations. The rate of workpiece erosion, and therefore the cutting efficiency, is very dependent on the material used. Aluminum oxide is an ideal abrasive for cutting brittle materials such as advanced ceramics and stones, yielding a twofold increase in cutting power as compared to garnet. However, aluminum oxide is not much more effective than garnet for cutting ductile materials. The abrasive size and flow rate also greatly affect the quality of the cut surface as well as the cutting efficiency. The range of mesh sizes commonly used for industrial applications is 50–120 (particle size 300–125 μm or 0.012–0.005 in.), and the range of abrasive flow rates is 0.5–1.5 lb/min (3.8–11.4 g/s). *See* ABRASIVE.

Standoff and traverse speed. The standoff distance of the nozzle tip from the workpiece ranges 0–0.2 in. (0–5 mm). A standoff in this range has a minimal influence on the cutting efficiency. The traverse speed of the cutting head is the most critical parameter. The depth of cutting as well as the surface finish is primarily governed by the traverse speed of the nozzle tip, which is controlled by mechanical manipulation. Therefore, the traverse speed is used as a single control parameter in most nonprecision cutting operations.

Thomas J. Kim

Bibliography. H. Louis et al., Investigation of human bones cut with high pressure water-jet, *Proceedings of 16th International Conference on Water Jetting*, October 2002; D. Miller, Micro abrasive water-jet cutting, *Proceedings of 2001 WJTA American Waterjet Conference*, August 2001; A. W. Momber (ed.), *Water Jet Applications in Construction Engineering*, 1998; A. W. Momber and R. Kovacevic, *Principles of Abrasive Water Jet Machining*, 1998; D. A. Summers, *Waterjetting Technology*, 1995; R. A. Tikhomirov et al., *High Pressure Jetting*, 1992.

Water pollution

A change in the chemical, physical, biological, and radiological quality of water that is injurious to its existing, intended, or potential uses. (for example, boating, waterskiing, swimming, the consumption of fish, and the health of aquatic organisms and ecosystems). The term "water pollution" generally refers to human-induced (anthropogenic) changes to water quality. Thus, the discharge of toxic chemicals from a pipe or the release of livestock waste into a nearby water body is considered pollution. Conversely, nutrients that originate from animals in the wild (for example, a herd of elk) or toxins that originate from natural processes (for example, red tides) are not considered pollution.

Contaminants. The contamination of ground water, rivers, lakes, wetlands, estuaries, and oceans can threaten the health of humans and aquatic life. Sources of water pollution are generally divided into two categories (**Table 1**). The first is point-source pollution, in which contaminants are discharged from a discrete location. Sewage outfalls and the 1989 Exxon *Valdez* oil spill are examples of point-source pollution. The second category is non-point-source or diffuse pollution, referring to all of the other discharges that deliver contaminants to water bodies. Acid rain and unconfined runoff from agricultural or urban areas are examples of non-point-source pollution. The principal contaminants of water are shown in (**Table 2**). These include toxic chemicals, nutrients and biodegradable organics, and bacterial and viral pathogens.

Point sources. Since the passage of the Clean Water Act in 1972, there has been considerable progress in reducing the amount of pollution originating from municipal and industrial point sources. Most wastewater treatment plants are designed to accomplish the removal of suspended solids, biodegradable organics, and pathogenic organisms. The impact of the discharge of biodegradable organics can be measured in terms of the buildup of sludge deposits and depletion of the dissolved oxygen resources of water bodies. This situation led to requirements for secondary treatment of wastewaters. Similarly, concern over the toxicity caused by the discharge of heavy metals in treated effluents led to the development of effective pretreatment programs.

Nonpoint sources. Reducing contamination from nonpoint sources has been considerably more difficult, in part because these inputs are widely distributed and highly variable. To better understand the relationship between water quality and chemical use, land use, climate geology, topography, and soils, Congress appropriated funds for the National Water-Quality Assessment (NAWQA) Program in 1991. The NAWQA Program is an ongoing investigation by the U.S. Geological Survey in river basins and aquifers around the country.

Transport and transformation processes. Contaminants discharged to water bodies are subject to a variety of transport and transformation processes and operations that can alter their composition. The physical, chemical, and biological processes that control the fate of the contaminants discharged to water bodies are numerous and varied. It is convenient to divide them into transport processes that affect all water-quality parameters in the same way and fate and transformation processes which are constituent-specific.

TABLE 1. Identification of point and non-point sources of contaminants discharged to water bodies*

Point sources	Non-point (diffuse) sources
Municipal and industrial wastewater effluents	Return flow from irrigated agriculture (specifically excluded from point source definition by Congress)
Runoff and leachate from solid waste disposal sites	Other agricultural and silvicultural runoff and infiltration from sources other than confined concentrated animal operations
Runoff and infiltrated water from concentrated animal feeding operations	Unconfined pastures of animals and runoff from range land
Runoff from industrial sites not connected to storm sewers	Urban runoff from sewered communities with a population of less than 100,000 not causing a significant water-quality problem
Storm sewer outfalls in urban centers with a population of more than 100,000	Urban runoff from unsewered areas
Combined sewer overflows	Runoff from small or scattered (less than 2 hectares or 5 acres) construction sites
Leachate from solid waste disposal sites	Septic tank surfacing in areas of failing septic tank systems and leaching of septic tank effluents
Runoff and drainage water from active mines, both surface and underground, and from oil fields	Wet and dry atmospheric deposition over a water surface (including acid rainfall)
Other sources, such as discharges from vessels, damaged storage tanks, and storage piles of chemicals	Flow from abandoned mines (surface and underground), including inactive roads,tailings, and spoil piles
Runoff from construction sites that are larger than 2 hectares (5 acres)	Activities on land that generate wastes and contaminants, such as deforestation and logging; wetland drainage and conversion; channeling of streams; building of levees, dams, causeways, and flow diversion facilities on navigable waters; construction, and development of land; interurban transportation; military training, maneuvers, and exercises; and mass outdoor recreation

*After V. Novotny and H. Olem, *Water Quality: Prevention, Identification, and Management of Diffuse Pollution*, Van Nostrand Reinhold, 1994.

After initial dilution, contaminants discharged to a water body are transported by two basic processes, advection and dispersion. Advection refers to the transport of a constituent resulting from the flow of the water in which the constituent is dissolved or suspended. Turbulent velocity fluctuations, in conjunction with concentration gradients and molecular diffusion, lead to a mass transport phenomenon called dispersion.

The principal fate and transformation processes that affect contaminants discharged to the environment are operative in most water bodies (**Table 3**). The relative importance of individual fate and transformation processes will be site-specific and will depend on the water-quality parameter under evaluation. For example, deoxygenation brought about by bacterial activity, surface reaeration, sediment oxygen demand, and photosynthesis and respiration are of major importance in assessing the oxygen resources of a stream.

Harmful effects on human health. Water pollution can threaten human health when pollutants enter the body via skin exposure or through the direct consumption of contaminated food or drinking water. For example, many states have issued fish consumption advisories following the detection of mercury in fish tissues. Other priority pollutants, including dichlorodiphenyl trichloroethane (DDT) and polychlorinated biphenyls (PCBs), persist in the natural environment and bioaccumulate in the tissues of aquatic organisms. These persistent organic pollutants are transferred up the food chain (in a process called biomagnification), and they can reach levels of concern in fish species that are eaten by humans. Finally, bacteria and viral pathogens can pose a public health risk for those who drink contaminated water or eat raw shellfish from polluted water bodies. *See* ENVIRONMENTAL TOXICOLOGY; FOOD WEB.

TABLE 2. Contaminants of concern found in point and non-point discharges to water bodies

Contaminants	Concern
Atmospheric pollutants	Acid rain leads to acid (low pH) conditions; deposition of potentially toxic constituents
Biodegradable organics	Depletion of natural oxygen resources and the development of septic conditions
Current-use pesticides	Toxicity to aquatic biota
Dissolved inorganics (such as total dissolved solids)	Inorganic constituents added by usage; reclamation and reuse applications
Heat	Growth of undesirable aquatic life; threat to sensitive life forms, upset of ecological balances
Heavy metals	Toxicity to aquatic biota; many metals are also classified as priority pollutants
Nutrients (nitrogen and phosphorus)	Growth of undesirable aquatic life; eutrophication
Pathogenic organisms	Communicable diseases
Priority organic pollutants	Suspected carcinogenicity, mutagenicity, teratogenicity, or high acute toxicity; many priority pollutants resist conventional treatment methods (known as refractory organics); long-term effects are unknown
Suspended solids (inorganic and organic)	Siltation; formation of sludge deposits and anaerobic conditions

TABLE 3. Fate and transformation processes affecting contaminants discharged to water bodies

Process	Constituents affected
Adsorption and desorption	Metals; trace organics; NH_4^+; PO_4^{-3}
Bacterial conversion, aerobic and anaerobic	Biochemical oxygen demand; nitrification; denitrification; sulfate reduction; anaerobic fermentation (in bottom sediments); conversion of priority organic pollutants
Chemical reactions (hydrolysis, ion exchange, oxidation-reduction, and so on)	Decomposition of organic compounds; specific ion exchange; element substitution
Filtration	Suspended matter; colloidal particles
Flocculation	Suspended matter; colloidal particles
Gas absorption and desorption	O_2; CO_2; CH_4; NH_3; H_2S
Natural decay	Plants; animals; protists (algae, fungi, protozoa); eubacteria (most bacteria); archaebacteria; viruses; radioactive substances
Photochemical reactions	Oxidation of inorganic and organic compounds
Photosynthesis and respiration	Algae, duckweed; submerged macrophites; NH_4^+; PO_4^{-3}; pH
Sedimentation	Suspended solids
Sediment oxygen demand	O_2, particulate biochemical oxygen demand
Surface reaeration	O_2; CO_2
Volatilization	Volatile organic compounds; NH_3; CH_4; H_2S, other gases

Harmful effects on aquatic species. Contaminants have a significant impact on aquatic ecosystems. for example, enrichment of water bodies with nutrients (principally nitrogen and phosphorus) can result in the growth of algae and other aquatic plants that shade or clog streams. If wastewater containing biodegradable organic matter is discharged into a stream with inadequate dissolved oxygen, the water downstream of the point of discharge (typically an outfall) will become anaerobic and will be turbid and dark. Settleable solids, if present, will be deposited on the streambed, and anaerobic decomposition will occur. Over the reach of stream where the dissolved-oxygen concentration is zero, a zone of putrefaction will occur with the production of hydrogen sulfide, ammonia, and other odorous gases. Because many fish species require a minimum of 4–5 mg of dissolved oxygen per liter of water, they will be unable to survive in this portion of the stream. In addition to reductions in dissolved oxygen, aquatic species are sensitive to changes in other physical habitat factors, including pH, temperature, and suspended solids.

Direct exposures to toxic chemicals is also a health concern for individual aquatic plants and animals. For example, pesticides are used to kill undesirable or nuisance organisms in many urban and agricultural areas. These chemicals are frequently transported to lakes and rivers via runoff, and they can have unintended and harmful effects on aquatic life. Obvious signs of contaminant exposure in fish from polluted environments include lesions, tumors, and skeletal deformities. Toxic chemicals have also been shown to reduce the growth, survival, reproductive output, and disease resistance of exposed organisms. These effects, while subtle, can have important consequences for the viability of aquatic populations and communities. *See* INSECTICIDE.

Effluent discharge. Wastewater discharges are most commonly controlled through effluent standards and discharge permits. In the United States, the National Pollution Discharge Elimination System (NPDES), administered by the individual states with federal Environmental Protection Agency (EPA) oversight, is used for the control of wastewater discharges. Under this system, discharge permits are issued with limits on the quantity and quality of effluents. These limits are based on a case-by-case evaluation of potential environmental impacts and, in the case of multiple dischargers, on waste load allocation studies aimed at distributing the available assimilative capacity of the water body. Discharge permits are designed as an enforcement tool, with the ultimate goal of meeting ambient water-quality standards.

Water-quality standards and criteria. Water-quality standards are sets of qualitative and quantitative criteria designed to maintain or enhance the quality of receiving waters. In the United States, these standards are promulgated by the individual states. Receiving waters are divided into several classes depending on their uses, existing or intended, with different sets of criteria designed to protect uses such as drinking water supply, bathing, boating, fresh-water and shellfish harvesting, and outdoor sports for seawater.

Toxicity studies. For toxic compounds, chemical-specific or whole-effluent toxicity studies are used to develop standards and criteria. In the chemical-specific approach, individual criteria are used for each toxic chemical detected in the wastewater. Criteria can be developed to protect aquatic life against acute and chronic effects and to safeguard humans against deleterious health effects, including cancer. The chemical-specific approach, however, does not consider the possible additive, antagonistic, or synergistic effects of multiple chemicals. The biological availability of the compound, which depends on its form in the wastewater, is also not considered in this approach.

The whole-effluent toxicity approach can be used to overcome the shortcomings of the chemical-specific approach. In the whole-effluent approach, toxicity or bioassay tests are used to determine the concentration at which the wastewater induces acute or chronic toxicity effects. In bioassay testing, selected organisms are exposed to effluent diluted in various ratios with samples of receiving water. At

various points during the test, the organisms affected by various effects, such as lower reproduction rates, reduced growth, or death, are quantified. To protect aquatic life discharge limits are established based on the results of the tests. *See* HAZARDOUS WASTE; SEWAGE TREATMENT.

Nathaniel Scholz; George Tchobanoglous

Bibliography. V. Novotny, *Water Quality: Diffuse Pollution and Watershed Management*, 2d ed., 2002; G. Tchobanoglous, F. L. Burton, and H. D. Stensel, *Wastewater Engineering: Treatment and Reuse*, 4th ed., 2002; G. Tchobanoglous and E. D. Schroeder, *Water Quality: Characteristics, Modeling, Modification*, 1985; K. M. Vigil, *Clean Water: An Introduction to Water Quality and Pollution Control*, 2003.

Water resources

The Earth's water supply and its natural distribution. Although water is a renewable resource, which is continually being replaced by precipitation, it is not evenly distributed and is scarce in many areas.

Water reservoirs. Water is stored on the Earth's surface in a number of places called reservoirs (**Table 1**).

Oceans. By far the largest reservoir is the ocean, which contains 96% of the Earth's water and occupies more than two-thirds of the Earth's surface. Ocean water, being saline, is not generally available for human consumption, although it can be used for some purposes, mainly thermoelectric power. *See* HYDROSPHERE; OCEAN.

Glaciers. Fresh water makes up only about 4% of the Earth's water. The largest freshwater reservoir is glacial ice, at 3%. Most of this ice (about 85%) occurs as continental glaciers in Antarctica and less than 10% in the Greenland ice sheet. Alpine or mountain glaciers, which occur in mountain valleys on the continents, contain a small part of the total ice. *See* HYDROLOGY; TERRESTRIAL WATER.

Ground water. The largest reservoir of available fresh water is ground water (1.05% of total water), which is stored in the pores and spaces in rocks, sand, gravel, and soil under the Earth's surface. The top plane of the ground water is referred to as the water table, below which all the spaces are filled with water. About half of the ground water occurs quite near the Earth's surface (<0.8 km) and this is an important source of water for human consumption. Although shallow ground water is continually being refilled by precipitation trickling down to the water table, the rate of recharge is very slow and often takes hundreds or thousands of years. This makes many ground-water aquifers a nonrenewable resource. The rest of the ground water, while at greater depths, does not occur much deeper than a few kilometers, where the pressure of the overlying rock becomes so great that pore space disappears. Deep ground water is harder to recover and is more likely to be saline. A smaller amount of water occurs in the soil above the water table, where both air and water fill the pore spaces; this water is referred to as soil moisture and is tightly held in the pores. *See* AQUIFER; ARTESIAN SYSTEM.

Lakes, rivers, and other reservoirs. Fresh-water lakes and rivers on the Earth's surface contain only 0.01% of the Earth's water. This water is generally available for human consumption. There is also an even smaller reservoir of water in the atmosphere (0.001%), where the water occurs as water vapor gas. The smallest reservoir of water occurs in the biosphere, within plants and animals (0.0006%).

To summarize, the main fresh-water resources available for humans on the Earth's surface are ground water and lake and river water, which together only constitute about 1.1% of the Earth's total water. *See* GROUND-WATER HYDROLOGY; SURFACE WATER.

Hydrologic cycle. Water does not permanently remain in any one reservoir on the Earth but is continually in motion through the hydrologic or water cycle (**Fig. 1**). Residence time is the length of time water spends in a particular reservoir before being removed through the water cycle. The ocean has a residence time of 36,000 years, the time it would take river flow to replace the water in the ocean. *See* ATMOSPHERIC GENERAL CIRCULATION.

The water cycle is driven by solar energy. Water is evaporated from the oceans to the atmosphere in the form of water vapor gas (a flux of 435,400 km^3/year). Part of this water vapor cools and is precipitated back to the oceans in the form of rain (398,000 km^3/year). Some water vapor (37,400 km^3/year) is transported through the atmosphere to the continents where it joins water vapor evaporated from the land (69,600 km^3/year). The vapor cools and precipitates out as rain or snow (107,000 km^3/year). Much of this precipitation runs along the Earth's surface to rivers, which flow to the oceans. A small amount of precipitation trickles down to join the ground water.

TABLE 1. World water*

Location	Surface area, km^2	Water volume, 10^3 km^3	Percent of total
World ocean	362,000,000	1,400,000	95.96
Mixed layer		50,000	
Thermocline		460,000	
Abyssal		890,000	
Glacial ice	18,000,000	43,400	2.97
Ground water		15,300	1.05
Lakes, fresh	855,000	125	0.009
Rivers		1.7	0.0001
Soil moisture		65	0.0045
Atmosphere†		15.5	0.001
Biosphere		2	0.0006
TOTAL	510,000,000	1,459,000	100

*SOURCE: E. K. Berner and R. A. Berner, *Global Environment: Water, Air and Geochemical Cycles*, Prentice Hall, 1996; National Research Council (NRC), *Global Change in the Geosphere—Biosphere*, National Academy Press, 1986; R. L. Nace, Terrestrial water, in AccessScience@McGraw-Hill, http://www.accessscience.com.DOI 10.1036/1097–8542.685800, last modified July 15, 2002 (surface area).

†As liquid water.

The flow of water in rivers to the ocean, which can be estimated reliably, is about 37,400 km^3/year. The water vapor flux from the oceans to the land is set equal to the river flux from the land to the oceans.

The total amount of water on the Earth's surface in the various reservoirs remains roughly constant over time. The general belief is that the amount of water on or near the Earth's surface has not changed greatly since 3.8 billion years ago. *See* SURFACE WATER.

Rivers. The rivers with the greatest volume of water discharged to the oceans are listed in **Table 2**. The first 13 rivers make up about 38% of the total water discharge to the oceans, which is approximately 37,400 km^3/yr. The Amazon River of Brazil and central South America alone comprises 17% of the total. There are three American rivers that are among the world's 17 largest rivers: the Mississippi, St. Lawrence, and Columbia rivers. *See* RIVER.

Size and distribution. The amount of river runoff from the continents is determined by the difference between precipitation and evaporation, or net precipitation in the river basin. Ultimately, the amount of precipitation is determined by the atmospheric circulation of water vapor. Evaporation rates depend upon temperature, and decrease with increasing distance from the equator. Most of the world's large rivers occur in two belts where precipitation exceeds evaporation, an equatorial belt from 10°N latitude to 10°S latitude and a temperate belt from 30 to 60°N and S latitude.

The two largest rivers, the Amazon and the Congo rivers, occur in the equatorial belt. The temperate belt has two major rivers, the Mississippi in the United States and the Yangtze in China. Although precipitation in the subarctic regions (60°N–70°N) is low, there is also low evaporation, resulting in net precipitation and a number of large rivers such as the Yenisei and the Lena in northern Russia and the Mackenzie in the Canadian Arctic. The 20–30°N and S belt, between the equatorial and temperate zones, has many deserts and few rivers because of descending dry air from the overall atmospheric circulation.

TABLE 2. Major world rivers by discharge to ocean

	River	Location	Discharge, km^3/yr
1	Amazon	South America	6300
2	Zaire (Congo)	Africa (Congo)	1250
3	Orinoco	South America	1100
4	Yangtze (Chiang)	Asia (China)	900
5	Bramaputra	Asia (India)	603
6	Mississippi	United States	580
7	Yenisei	Asia (Russia)	560
8	Lena	Asia (Russia)	525
9	Mekong	Asia (Vietnam)	470
10	Ganges	Asia (India)	450
11	St. Lawrence	North America	447
12	Parana	South America	429
13	Irrawaddy	Asia (Burma)	428
15	Mackenzie	North America (Canada)	306
17	Columbia	United States	251
	TOTAL all rivers		37,400

SOURCE: E. K. Berner and R. A. Berner, *Global Environment: Water, Air, and Geochemical Cycles*, Prentice Hall, Upper Saddle River, NJ, 1996.

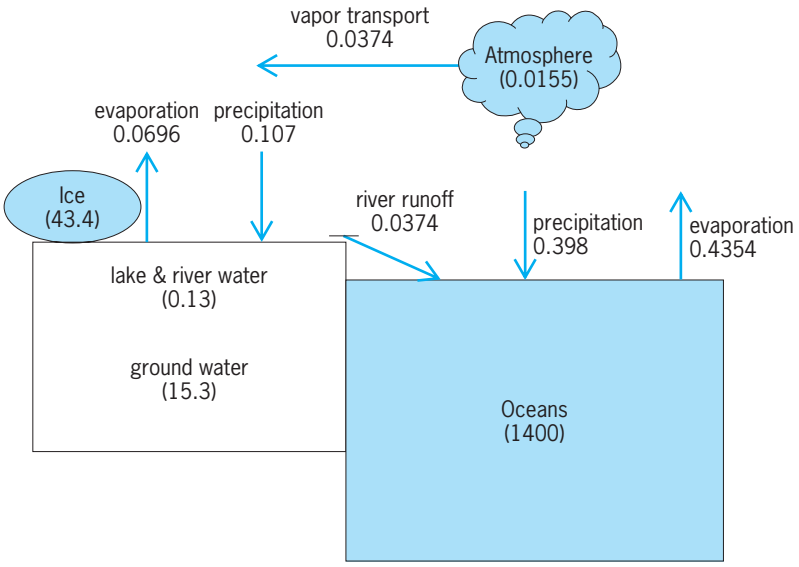


Fig. 1. Water cycle. Volumes of reservoirs in parentheses (in 10^6 km^3); fluxes in 10^6 km^3/yr. (*Data source: E. K. Berner and Robert A. Berner, Global Environment: Water, Air, and Geochemical Cycles, Prentice Hall, Upper Saddle River, NJ, 1996*)

Usage problems. The result of the uneven distribution of net precipitation and world rivers is that many areas do not have adequate water resources. For example, large parts of the flow of the Amazon and Congo rivers, which account for 18.5% of the total global river flow, are geographically inaccessible to populated areas. The remote northern rivers of North America and Eurasia (Mackenzie, Lena, and Yenisei) are also inaccessible. In total, about 20% of world river flow is geographically inaccessible to populated areas and thus not available for human use.

The western part of the United States is semiarid to arid, with precipitation generally declining as one moves west from the Mississippi River. The Colorado River, for example, flows through a generally arid region. Water is allocated by law to various states and withdrawn from the Colorado along its U.S. route until there is almost no flow by the time the river reaches Mexico. The water that remains contains high concentrations of dissolved minerals because of irrigation use and evaporation from large reservoirs behind dams, and therefore Mexico's share of the Colorado River must be desalinated. There are disputes over Colorado River water between California, Arizona, Nevada, and Mexico. The northwestern United States is an exception to the generally dry western United States, with a major river, the Columbia, fed by heavy precipitation in the coastal mountain ranges.

Rivers in other dry parts of the world, such as the Nile in Egypt, have had their flow greatly reduced due to dams and irrigation. At times, the Nile is reduced to zero flow.

Lakes. Freshwater lakes only make up 0.009% of the worlds water by volume (125,000 km^3), but they are important water resources. The world's major freshwater lakes are listed in **Table 3** in order of surface area.

TABLE 3. Major world freshwater lakes

	Lake	Location	Area, km^2	Maximum depth, m	Length, km
1	Superior	United States–Canada	82,103	406	563
2	Victoria	Africa	69,485	82	402
3	Huron	United States–Canada	59,570	229	331
4	Michigan	United States	57,757	281	494
5	Tanganyika	Africa	32,893	1470	676
6	Baikal	Russia	31,500	1620	636
7	Great Bear	Canada	31,329	446	306
8	Nyasa	Africa	28,879	695	579
9	Great Slave	Canada	28,570	614	480
10	Erie	United States–Canada	25,667	64	388
11	Winnipeg	Canada	24,390	18	428
12	Ontario	United States–Canada	19,011	244	311

SOURCE: U.S. Geological Survey.

Size and distribution. The American-Canadian Great Lakes (Superior, Michigan, Huron, Erie, and Ontario) are among the world's major lakes. They contain approximately 25,500 km^3 of water, or one-fifth of the world's fresh water, and are connected with the water finally flowing eastward from Lake Ontario into the St. Lawrence River and on to the Atlantic. The Great Lakes were formed when the last Pleistocene glacier melted about 15,000 years ago and owe their origin to glacial features: glacial erosion, deposits of sand and gravel left by the melting ice, and the general downwarping of the Earth's crust by the weight of glacial ice in north central North America. In addition to the Great Lakes, there are large numbers of small glacial lakes in Canada, Minnesota, Michigan, western New York, and New England. The Finger Lakes in upstate New York are conspicuous examples of glacial erosion. Two other large glacial lakes and many smaller lakes occur in northwestern Canada, including Great Bear Lake and Great Slave Lake, and Lake Winnipeg in central Canada. *See* LAKE.

Lake Baikal in Russian Asia is the world's largest lake by volume (23,000 km^3) because it is also the deepest at 1620 m. It alone contains about one-fifth of the Earth's fresh lake water. Lake Baikal occurs in a rift valley, a narrow downdropped valley with steep faults on either side. This type of geologic setting contributes to its great depth. A number of large, deep African lakes, such as Lake Tanganyika and Lake Nyasa, are also formed in rift valleys. Lake Tanganyika, which is almost as deep as Lake Baikal, also contains a large volume of freshwater at 20,850 km^3.

Lake Chad in central Africa used to be one of the world's largest freshwater lakes by surface area although it was very shallow (5–8 m). However, a combination of extensive use for irrigation and long-term drought have reduced its size by 90% from 25,750 km^2 in the 1960s to 2150 km^2 in 2001.

There are several very large saline lakes. In fact, the world's largest lake, the Caspian Sea (at 371,000 km^2) at the border between Asia and Europe, is saline. It was named a "sea" because of its salinity. The Aral Sea is another large saline lake, although it has been shrunk extensively by the use of its water for irrigation; its original size was 67,340 km^2 and it is now about half that large. In addition, the salty beds left behind the dried-up lake are exposed to the wind, resulting in dust storms that cause severe air pollution. The largest American saline lake, the Great Salt Lake in Utah, which is more saline than seawater, is the remnant of an originally freshwater glacier lake.

Usage problems. One major problem with freshwater lake quality in the United States has been contamination of lakes with nitrogen and phosphorus. This causes overgrowth of algae and green murky water as well as loss of oxygen in the bottom water when the algae die and decay. Phosphorus has been removed from detergents, improving lake water quality, but there remains an ongoing problem with phosphate and nitrogen from sewage effluent and lawn and agricultural runoff.

Acid lakes, particularly in the northeastern United States and in Scandinavia, result from the production of air pollution upwind from the lakes. Sulfuric acid pollution comes from coal burning power plants and nitric acid from automotive pollution and power plants. The air pollution is carried by the atmospheric circulation to susceptible lakes and then falls as acid rain.

United States water use. Freshwater use in the United States, which totaled 345 billion gallons/day in 2000, is shown in **Fig. 2**. The two largest uses, at about 40% each, are irrigation of crops and thermoelectric power generation. Public water supply accounts for 12%, followed by industry at 6%. The

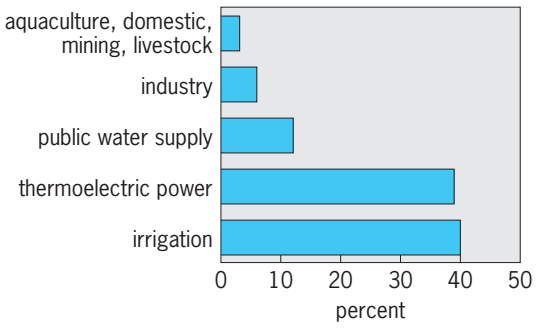


Fig. 2. Freshwater use in the United States in 2000. (*Data source: U.S. Geological Survey Cir. 1268, 2004*)

last four uses are less than 1% each: aquaculture, domestic (well) water, mining, and livestock.

The sources of freshwater are surface water (76%) and ground water at (24%). The largest use of ground water is for irrigation, which has increased to 42% of the total water use or 58 billion gallons/day. Ground water also makes up 37% of the public water supply. These two uses make up most of the total groundwater withdrawal.

The largest withdrawal of ground water occurs in three states, California, Texas and Nebraska, where ground water is used primarily for irrigation. The Ogallala Aquifer, which lies under the High Plains states and is a big ground-water source, was drawn down by more than 100 ft from 1940 to 1980 in southwest Kansas, Texas, and the Oklahoma panhandle. However, better conservation methods have since been implemented, such as more efficient irrigation to reduce water loss.

Worldwide water use. The uses of water worldwide are 70% for agriculture, 10% for domestic purposes such as drinking water, and 20% for industry (more than half of which is used for hydropower). Countries that have scarce water include a number in the belt of low precipitation (20–30° N and S latitude) such as the northern tier of Africa (Mauritania, Algeria, Morocco, Libya, Niger, and Egypt) and the Middle East (Saudi Arabia, Palestine, Syria, and Jordan). Worldwide there are 500 million people in countries with scarce water.

Water for human consumption is unsafe in many places, particularly in the developing countries. It is estimated that as much as 80% of diseases in developing countries are water-related, and 1.7 million people, often children, die from these diseases mainly in Africa and southeast Asia. Typical diseases are diarrhea, cholera, typhoid, and malaria. The main problem is that unsafe disposal of human and animal waste contaminates water for domestic use and irrigation.

More than 50% of the water used by industry (20% of the total) is used for hydropower plants. These plants provide one-fifth of the world's electricity. Hydropower is relatively clean and nonpolluting and is renewable. Dams used for hydropower generation also store water resources for agricultural irrigation, flood prevention, and domestic use. The downside of dams is their displacement of people living in the area of the reservoir. Industrial uses of water can lead to pollution of rivers and aquifers by heavy metals (such as mercury and lead) and persistent organic pollutants (such as DDT and PCB's from electrical insulation).

Agriculture uses 70% of water worldwide, primarily for irrigation. About 65% of irrigation water is "consumed" in distribution and application and by crops and not available for reuse. Irrigation can be wasteful of water and can lead to salt buildup in soils if the soil is poorly drained. Agricultural and lawn runoff often cause overfertilization of water from nitrate and phosphate, causing algal blooms and loss of oxygen in bottom water of rivers, lakes, and estuaries. There have also been problems with agricultural pesticides polluting ground and surface water.

Desalinization of salt water currently supplies only about 0.1% of fresh water. It is expensive since it requires a lot of energy. Thus, it is used primarily for drinking water in water-poor areas. *See* WATER DESALINATION.

Elizabeth K. Berner

Bibliography. E. K. Berner and R. A. Berner, *Global Environment: Water, Air and Geochemical Cycles*, Prentice Hall, Upper Saddle River, NJ, 1996; R. Clarke and J. King, *The Water Atlas*, New Press, New York, 2004; P. H. Gleick, *The World's Water 2004–2005: The Biennial Report on Freshwater Resources (World's Water)*, Island Press, Washington, DC, 2005; P. H. Gleick, Making every drop count, *Sci. Amer.*, 284(2):40–45, 2001; S. S. Hutson et al., Estimated use of water in the United States in 2000, U.S. Geological Survey Circ. 1268, 2004; R. L. Nace, Terrestrial water, in AccessScience@ McGraw-Hill, http://www.accessscience.com.DOI 10.1036/1097-8542.685800, last modified July 15, 2002; S. L. Postel, G. C. Daly, and P. R. Ehrlich, Human appropriation of renewable fresh water, *Science*, 271:785–88, Feb. 9, 1996.

Water softening

The process of removing divalent cations, usually calcium or magnesium, from water. When a sample of water contains more than 120 mg of these ions per liter (0.016 oz/gal), expressed in terms of calcium carbonate ($CaCO_3$), it is generally classified as a hard water. Hard waters are frequently unsuitable for many industrial and domestic purposes because of their soap-destroying power and tendency to form scale in equipment such as boilers, pipelines, and engine jackets. Therefore it is necessary to treat the water either to remove or to alter the constituents for it to be fit for the proposed use.

The principal water-softening processes are precipitation, cation exchange, electrical methods, or combinations of these. The factors to be considered in the choice of a softening process include the raw-water quality, the end use of softened water, the cost of softening chemicals, and the ways and costs of disposing of waste streams.

Lime-soda processes. The principal reactions (1)–(7) involved in the lime-soda processes are the

$$CO_2 + Ca(OH)_2 = CaCO_3 \downarrow + H_2O \quad (1)$$

$$Ca(HCO_3)_2 + Ca(OH)_2 = 2CaCO_3 \downarrow + 2H_2O \quad (2)$$

$$Mg(HCO_3)_2 + Ca(OH)_2 = CaCO_3 \downarrow + MgCO_3 + 2H_2O \quad (3)$$

$$MgCO_3 + Ca(OH)_2 = CaCO_3 \downarrow + Mg(OH)_2 \quad (4)$$

$$2NaHCO_3 + Ca(OH)_2 = CaCO_3 \downarrow + Na_2CO_3 + 2H_2O \quad (5)$$

$$MgSO_4 + Ca(OH)_2 = Mg(OH)_2 \downarrow + CaSO_4 \quad (6)$$

$$CaSO_4 + Na_2CO_3 = CaCO_3 \downarrow + Na_2SO_4 \quad (7)$$

precipitation of bicarbonates of calcium and magnesium through the addition of lime, and precipitation of noncarbonates (sulfates, chlorides, and nitrates of magnesium and calcium) by reacting with soda ash.

In reactions (1) and (5), carbon dioxide (CO_2) and sodium bicarbonate ($NaHCO_3$) are not part of the hardness but will consume lime [$Ca(OH)_2$] and therefore must be considered in calculating the required lime amount. Reactions (2), (3), and (4) indicate the removal of carbonate hardness, while reactions (6) and (7) show the removal of magnesium and calcium noncarbonate hardness ($MgSO_4$ and $CaSO_4$), respectively. It is evident from these equations that the amounts of lime and soda ash (Na_2CO_3) required to soften a water may be calculated from the concentrations of free carbon dioxide, bicarbonate, magnesium, and noncarbonate hardnesses.

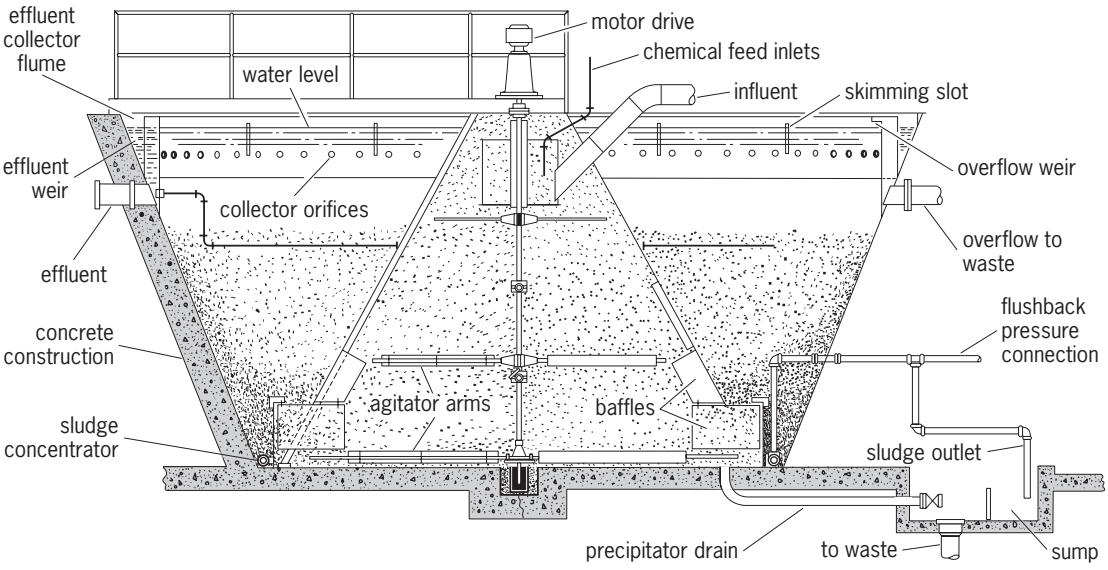
The types of equipment required for lime–soda ash softening processes generally includes rapid-mixing, flocculation, and sedimentation basins; chemical handling and feeding equipment; conventional rapid sand filters; and, sometimes, carbon dioxide–generating equipment. The conventional rapid-mixing, flocculation, and sedimentation basins are used to carry out the softening process. The chemical handling and feeding equipment includes the lime feeder, slaker, filters, and sludge-handling pumps. A typical precipitator type of water softening plant is shown in the **illustration**. The usual acceptable level of total hardness is 70–120 mg/liter (0.0093–0.016 oz/gal), expressed as $CaCO_3$. As soda ash is more costly than lime, it is advantageous to decrease the carbonate hardness to its lowest practical level. This would allow the maximum concentration of noncarbonate to remain and thus reduce softening costs. Moreover, it would require approximately 190% more soda ash than lime to remove the same amount of hardness.

Complete removal of calcium and magnesium hardnesses is not possible due to the slight solubility of calcium carbonate and magnesium hydroxide. Although, theoretically, it is feasible to reduce the hardness of water to about 25 ppm, in actual practice the range is up to 50–60 ppm or higher.

Variations in the lime-soda process. The fact that satisfactory hardness reduction may still cause deposit formation stimulated the development of methods designed to prevent scale formation by improving the stability of softened water. These methods include (1) conventional lime–soda ash softening followed by recarbonation, which makes it possible to reduce the calcium hardness to 15–20 mg/liter (0.0020–0.0027 oz/gal); (2) excess-lime treatment, a method most appropriate for waters containing 40 mg/liter (0.005 oz/gal) or more magnesium hardness (as $CaCO_3$) and usually requiring neutralization of the excess lime remaining in the solution either by recarbonation or by the use of bypassed raw water (split treatment) after treatment; (3) the hot-lime process, which takes advantage of the high reaction rates and rapid settling of the precipitates due to the decreased water viscosity at high temperatures (200–212°F or 93–100°C), with exhausted steam supplemented by live steam generally used for heating; (4) the combined lime-soda and zeolite process, a method claimed to be most suitable for water containing a high degree of temporary hardness, as water so produced contains a small amount of sodium bicarbonate.

Disposal of the large quantities of lime-softening sludge is a serious environmental problem and must be well planned in advance. The principal methods for disposal are discharge into water streams; discharge into sewers; vacuum filtration to form cake for landfill or as agricultural lime; ponding or lagooning; and recalcination of the dried sludge. The specific disposal method depends, among other things, on local regulations, effects on environment, and costs.

Ion-exchange softening. The basic principle involved in using ion-exchange processes for water



Cold lime-soda water softener, the sludge-blanket type with vertical concrete or steel precipitator.

softening is the utilization of the ability of certain insoluble substances to exchange cations with other cations dissolved in water. They are most appropriate for waters with high noncarbonate hardness or high magnesium content as the use of lime or lime-soda may be quite expensive in these cases. The exchangers used in the processes are zeolites and resins. One of the major advantages in using ion-exchange softening is that the process is continuous and the exchanger can be regenerated. Because the exchange reaction is reversible, after the cations in the exchanger are exhausted, the exchanger can be regenerated with a solution of sodium chloride. *See* ION EXCHANGE.

Zeolite process. Zeolites are crystalline alumino silicates which display cation-exchange properties. In a typical zeolite-softening plant, the hard water enters the top of a zeolite bed and filters down through the bed. Soft water is drawn off from the bottom. Regeneration is accomplished by flowing brine through the bed to remove calcium, and other noncarbonate hardnesses from the zeolite. Washing is generally done by passing water upward to remove traces of brine. The advantages of zeolite processes include production of low water hardness, application to raw water of variable hardness, operation under pressure, and low costs. *See* ZEOLITE.

Cation-exchange resins. The most commonly used cation-exchange resins are sulfonated copolymers of styrene and divinylbenzene. Raw water is generally charged downward through a resin bed which is designed to operate either with pressure or with gravity flow. Regeneration of the resin is accomplished by passing brine uniformly through the bed at a controlled rate through a distributor system. Problems associated with using cation-exchange resins are fouling of the exchanger by the presence of iron, manganese, and aluminum; loss of exchange capacity due to turbidity and bacterial slimes; and swelling caused by the presence of oxidizing agents. *See* WATER TREATMENT.

Yi Hua Ma

Bibliography. American Water Works Association, Inc., *Water Quality and Treatment*, 5th ed., 1999; M. W. Harza, *Water Treatment: Principles and Design*, 2d ed., 2005; A. P. Sincero and G. A. Sincero, *Physical-Chemical Treatment of Water and Wastewter*, 2002; Nalco Chemical Co., *The Nalco Water Handbook*, 2d ed., 1988.

Water supply engineering

A branch of civil engineering concerned with the development of sources of supply, transmission, distribution, and treatment of water. The term is used most frequently for municipal water works, but applies also to water systems for industry, irrigation, water reuse, and other purposes.

Sources of Water Supply

Underground waters, rivers, lakes, and reservoirs, the primary sources of freshwater, are replenished by rainfall. Some of this water flows to the sea through surface and underground channels, some is taken up by vegetation, and some is lost by evaporation. Freshwater may also be extracted from saline or brackish waters. Water not intended for potable use, such as for irrigation, ground-water replenishment, salt-water intrusion barriers, or habitat restoration, can be supplied by reusing highly treated wastewater.

Ground water. Water obtained from subsurface sources, such as sand and gravel and porous or fractured rocks, is called ground water. Ground water flows toward points of discharge in river valleys and, in some areas, along the seacoast. The flow takes place in water-bearing strata known as aquifers. The velocity may be a few feet to several miles per year, depending upon the permeability of the aquifer and the hydraulic gradient or slope. A steep gradient or slope indicates relatively high pressure, or head, forcing the water through the aquifer. When the gradient is flat, the pressure forcing the water is small. When the velocity is extremely low, the water is likely to be highly mineralized. If there is no movement, the water is rarely fit for use. *See* AQUIFER; GROUNDWATER HYDROLOGY.

Permeability is a measure of the ease with which water flows through an aquifer. Coarse sands and gravels, and limestone with large solution passages, have high permeability. Fine sand, clay, silt, and dense rocks (unless badly fractured) have low permeability.

Water table. In an unconfined stratum, the water table is the top or surface of the ground water. It may be within a few inches of the ground surface or hundreds of feet below. Normally it follows the topography. Aquifers confined between impervious strata may carry water under pressure. If a well is sunk into such an aquifer and the pressure is sufficient, water may be forced to the surface, resulting in an artesian well. The water-table elevation and artesian pressure may vary substantially with the seasons, depending upon the amount of rainfall recharging the aquifer and the amount of water taken from the aquifer. If pumpage exceeds recharge for an extended period, the aquifer is depleted and the water supply lost. *See* WATER TABLE.

Salt-water intrusion. Normally the ground-water flow is toward the sea. This normal flow may be reversed by overpumping and lowering of the water table or artesian pressure in an aquifer. Salt water flowing into the fresh-water aquifer being pumped is called salt-water intrusion.

Springs. Springs occur at the base of sloping ground or in depressions where the surface elevation is below the water table, or below the hydraulic gradient in an artesian aquifer from which the water can escape. Artesian springs are fed through cracks in the overburden or through other natural channels extending from the aquifer under pressure to the surface. *See* SPRING (HYDROLOGY).

Wells. Wells are vertical openings, excavated or drilled, from the ground surface to a water-bearing stratum or aquifer. Pumping a well lowers the water level in it, which in turn forces water to flow from the

aquifer. Thick, permeable aquifers may yield several million gallons daily with a drawdown (lowering) of only a few feet. Thin aquifers, or impermeable aquifers, may require several times as much drawdown for the same yields, and frequently yield only small supplies.

Dug wells, several feet in diameter, are frequently used to reach shallow aquifers, particularly for small domestic and farm supplies. They furnish small quantities of water, even if the soils penetrated are relatively impervious. Large-capacity dug wells or caisson wells, in coarse sand and gravel, are used frequently for municipal supplies. Drilled wells are sometimes several thousand feet deep.

The portion of a well above the aquifer is lined with concrete, stone, or steel casing, except where the well is through rock that stands without support. The portion of the well in the aquifer is built with open-joint masonry or screens to admit the water into the well. Metal screens, made of perforated sheets or of wire wound around supporting ribs, are used most frequently. The screens are galvanized iron, bronze, or stainless steel, depending upon the corrosiveness of the water and the expected life of the well. Plastic screens are sometimes used.

The distance between wells must be sufficient to avoid harmful interference when the wells are pumped. In general, economical well spacing varies directly with the quantity of water to be pumped and inversely with the permeability and thickness of the aquifer. It may range from a few feet to a mile or more.

Infiltration galleries are shafts or passages extending horizontally through an aquifer to intercept the ground water. They are equivalent to a row of closely spaced wells and are most successful in thin aquifers along the shore of rivers, at depths of less than 75 ft (23 m). These systems are sometimes called bank filtration because the river water is filtered in a natural or engineered system before it reaches the collector. The galleries are built in open cuts or by tunneling, usually with perforated or porous liners to screen out the aquifer material and to support the overburden.

Ranney wells consist of a center caisson with horizontal, perforated pipes extending radially into the aquifer. They are particularly applicable to the development of thin aquifers at shallow depths, and sometimes fed by river water from the surface.

Specially designed pumps, of small diameter to fit inside well casings, are used in all well installations, except in flowing artesian wells or where the water level in the well is high enough for direct suction lift by a pump on the surface (about 15 ft or 5 m maximum). Well pumps are set some distance below the water level, so that they are submerged even after the drawdown is established. Well-pump settings of 100 ft (30 m) are common, and they may exceed 300 ft (90 m) where the ground-water level is low. Multiple-stage centrifugal pumps are used most generally. They are driven by motors at the surface through vertical shafts, or by waterproof motors attached directly below the pumps. Wells are sometimes pumped by air lift, that is, by injecting compressed air through a pipe to the bottom of the well. *See* ARTESIAN SYSTEMS; WELL.

Surface water. Natural sources, such as rivers and lakes, and impounding reservoirs are sources of surface water. *See* DAM; RESERVOIR; SURFACE WATER.

Water is withdrawn from rivers, lakes, and reservoirs through intakes. The simplest intakes are pipes extending from the shore into deep water, with or without a simple crib and screen over the outer end. Intakes for large municipal supplies may consist of large conduits or tunnels extending to elaborate cribs of wood or masonry containing screens, gates, and operating mechanisms. Intakes in reservoirs are frequently built as integral parts of the dam and may have multiple ports at several levels to permit selection of the best water, which may vary seasonally. The location of intakes in rivers and lakes must take into consideration water quality, depth of water, likelihood of freezing, and possible interference with navigation. Reservoir intakes are usually designed for gravity flow through the dam or its abutments.

Wastewater reuse. Because of increasing demands on freshwater sources, it is necessary to use alternative water sources. Wastewater, or treated municipal sewage, is a water resource that can be effectively used in some cases. When sewage is treated by advanced methods, such as coagulation, filtration, membranes, or oxidation, the resulting water quality is often as good as or better than the water source that would normally receive it. This improved effluent can be used for nonpotable applications such as irrigating golf courses and agriculture, replenishing ground-water supplies and salt-water intrusion barriers, or restoring habitats. Federal or state guidelines ensure the safety of this water for its intended reuse. In some areas of the world, highly treated sewage effluent is effectively used as a direct supply at drinking water treatment plants. *See* WASTEWATER REUSE.

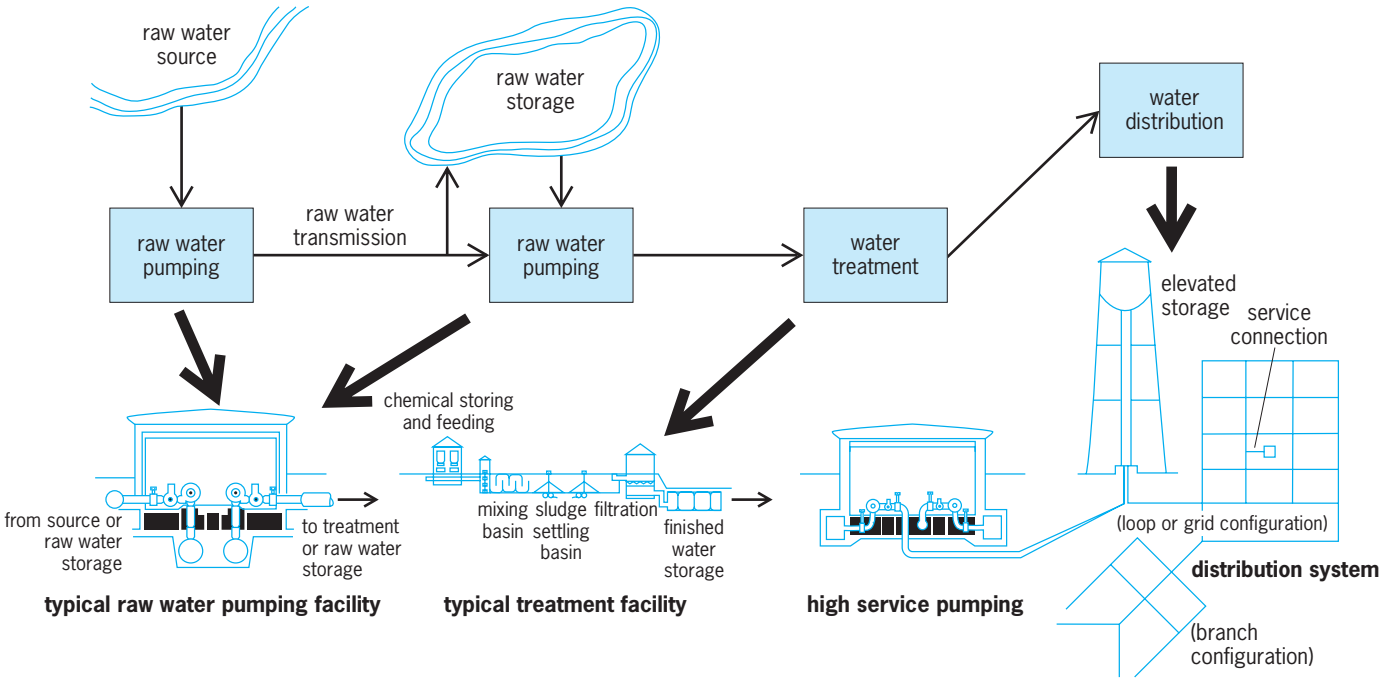
Water desalination. In many arid coastal communities, the demand for freshwater is outstripping the economically available supplies. One source of freshwater is treated saline or brackish water from the ocean or an estuary. Water desalination is typically achieved by treating the saline water through a reverse-osmosis process to separate salt ions from the water and recover the freshwater. This process requires a lot of energy and results in a portion of the water going to waste containing concentrated ions and pollutants. *See* WATER DESALINATION.

Transmission and Distribution

The water from the source must be transmitted to the community or area to be served and distributed to the individual customers after proper treatment.

Transmission mains. The major supply conduits, or feeders, from the source to the distribution system are called mains or aqueducts.

Canals. The oldest and simplest types of aqueducts, especially for transmitting large quantities of water, are canals. Canals are used where they can be built economically to follow the hydraulic gradient or slope of the flowing water and are open



Components of a water utility. (*M. J. Cullinane, Methodologies for the Evaluation of Water Distribution Systems Reliability/Availability, Ph.D. dissertation, University of Texas at Austin, May 1989*).

to the atmosphere. If the soil is suitable, the canals are excavated with sloping sides and are not lined. Otherwise, concrete or asphalt linings are used. Gravity canals are carried across streams or other low places by wooden or steel flumes, or under the streams by pressure pipes known as inverted siphons. *See* CANAL.

Tunnels. Used to transmit water through ridges or hills, tunnels may follow the hydraulic grade line and flow by gravity or may be built below the grade line to operate under considerable pressure. Rock tunnels may be lined to prevent the overburden from collapsing, to prevent leakage, or to reduce friction losses by providing a smooth interior. *See* TUNNEL.

Pipelines. Pipelines are a common type of transmission main, especially for moderate supplies not requiring large aqueducts or canals. Pipes are made of cast iron, ductile iron, steel, reinforced concrete, cement-asbestos, or wood. Pipeline material is determined by cost, durability, ease of installation and maintenance, and resistance to corrosion. The pipeline must be large enough to deliver the required amount of water and strong enough to withstand the maximum gravity or pumping pressure. Pipelines are usually buried in the ground for protection and coolness. *See* PIPELINE.

Distribution system. Included in the distribution system are the network of smaller mains branching off from the transmission mains, the house services and meters, the fire hydrants, and the distribution storage reservoirs. The network is composed of transmission or feeder mains, usually 12 in. (30 cm) or more in diameter, and lateral mains along each street, or in some cities along alleys between the streets. The mains are installed in grids so that lateral mains can be fed from both ends where possible. Mains fed from one direction only are called dead ends; they are less reliable and do not furnish as much water for fire protection as do mains with the grid. Valves at intersections of mains permit a leaking or damaged section of pipe to be shut off with minimum interruption of water service to adjacent areas. Water distributed for potable use typically originates at a water treatment facility (see **illustration**).

House services. Called house services, the small pipes are usually of iron, copper, or plastic extending from the water main in the street to the customer's meter at the curb line or in the cellar. In many cities, each service is metered, and the customer's bill is based on the water actually used.

Fire hydrants. Fire hydrants have a vertical barrel extending to the depth of the water main, a quick-opening valve with an operating nut at the top, and connections threaded to receive a fire hose. Hydrants must be reliable, and they must drain upon closing to prevent freezing.

Distribution reservoirs. These are used to supplement the source of supply and transmission system during peak demands, and to provide water during a temporary failure of the supply system. In small waterworks, the reservoirs usually equal at least one day's water consumption. In larger systems, the reservoirs are relatively smaller but adequate to meet fire-fighting demands. Ground storage reservoirs, elevated tanks, and standpipes are used as distribution reservoirs.

Ground storage reservoirs, if on high ground, can feed the distribution system by gravity, but otherwise it is necessary to pump water from the reservoir into the distribution system. Circular steel tanks and basins built of earth embankments, concrete, or rock masonry are used. Earthen reservoirs are

usually lined to prevent leakage or entrance of dirty water. The reservoirs should be covered to protect the water from dust, rubbish, bird droppings, and the growth of algae, but many older reservoirs without covers are in use, with some in large cities.

Elevated storage reservoirs are tanks on towers, or high cylindrical standpipes resting on the ground. Storage reservoirs are built high enough so that the reservoir will maintain adequate pressure in the distribution system at all times.

Elevated tanks are usually made of steel plate, mounted on steel towers. Wood is sometimes used for industrial and temporary installations. Standpipes are made of steel plate, strong enough to withstand the pressure of the water column. The required capacity of a standpipe is greater than that of an elevated tank because only the upper portion of a standpipe is sufficiently elevated for normal use.

Distribution-system design. To ensure the proper location and size of feeder mains and laterals to meet normal and peak water demands, a distribution system must be expertly designed. As the water flows from the source of supply or distribution reservoir across a city, the water pressure in the pipes is lowered by friction. The pressures required for adequate service depend upon the height of buildings, need for fire protection, and other factors, but 40 lb/in.^{2} (275 kilopascals) is the minimum pressure needed for good service. Higher pressures for fire fighting are obtained by booster pumps on fire engines that take water from fire hydrants. In small towns, adequate hydrant flows are the controlling factor in determining water-main size. In larger communities, the peak demands for air conditioning and lawn sprinkling during the summer months control the size of water main needed. The capacity of a distribution system is usually determined by opening fire hydrants and measuring simultaneously the discharge and the pressure drop in the system. The performance of the system when delivering more or less water than during the test can be computed from the pressure drops recorded.

An important factor in the economical operation of municipal water supplies is the quantity of water lost from distribution because of leaky joints, cracked water mains, and services abandoned but not properly shut off. Unaccounted-for water, including unavoidable slippage of customers' meters, may range from 10% in extremely well- managed systems to 30–40% in poor systems. The quantities flowing in feeder mains, the friction losses, and the amount of leakage are frequently measured by means of pitometer surveys. A pitometer is a portable meter that can be inserted in a water main under pressure to measure the velocity of flow, and thus the quantity of flow. *See* FLOW MEASUREMENT.

Pumping stations. Pumps are required wherever the source of supply is not high enough in elevation to provide gravity flow and adequate pressure in the distribution system. The pumps are high or low head depending upon the topography and pressures required. Booster pumps are installed on pipelines to increase the pressure and discharge, and adjacent to ground storage tanks for pumping water into distribution systems. Pumping stations usually include two or more pumps, each of sufficient capacity to meet demands when one unit is down for repairs or maintenance. The station must also include piping and valves arranged so that a break can be isolated quickly without cutting the whole station out of service.

Centrifugal pumps have replaced steam-driven reciprocating pumps in modern practice, although many of the old units continue to give good service. The centrifugal pumps are driven by electric motors, steam turbines, or diesel engines, with gasoline engines frequently used for standby service. The centrifugal pumps that are used most commonly are designed so that the quantity of water delivered decreases as the pumping head or lift increases. *See* CENTRIFUGAL PUMP; PUMP.

Automatic control of pumping stations is provided to adjust pump operations to variations in water demand. The controls start and stop pumps of different capacity as required. In the event of mishap or failure of a unit, alarms are sounded. The controls are activated by the water level in a reservoir or tank, the pressure in a water main, or the rate of the flow through a meter. A remote control system for pumps is often used, with the signals being transmitted over telephone wires.

Karl G. Linden; Richard Hazen

Water Treatment

Drinking water comes from surface and groundwater sources. Surface waters normally contain suspended matter, pathogenic organisms, and dissolved organic substances. Ground water normally contains dissolved minerals and gases. Both require treatment. Conventional water treatment processes include pretreatment, aeration, rapid mix, coagulation and flocculation, sedimentation, filtration, disinfection, and other unit processes to meet specific requirements (see illustration).

Aeration (air or oxygen into water) and air stripping (water into air) primarily are used to remove dissolved gases, such as hydrogen sulfide, which cause taste and odor, as well as to oxidize iron and manganese.

Rapid mix. Rapid (flash) mix is the unit operation where a coagulant chemical, such as alum, is applied and uniformly dispersed. Typically, the tank is equipped with a mechanical mixer and sized for a detention time of 30–60 s.

Coagulation and flocculation. The chemical process (coagulation) of neutralizing the chemical and electrostatic forces that repel particles is followed by the physical process (flocculation) of encouraging particle collisions and agglomerating particles for successful removal by gravity sedimentation. Conventional systems include parallel, rectangular basins equipped with slow-speed mixers to stimulate flocculation. Flocculation time, typically not less than 30 min, is a function of mixing intensity and coagulant dosage.

Sedimentation. Settleable particulate matter, chemical floc, and chemical precipitates are

removed by gravity settling. A typical facility has multiple basins with a working detention time of 2–4 h. Inlet and outlet arrangements, velocity through the basin, hydraulic and solids loading rates, and sludge storage and removal are important design considerations. *See* SEDIMENTATION (INDUSTRY).

Filtration. Using a porous media, the filtration process passes water while capturing colloidal and suspended solids not removed during sedimentation. Typical filter designs include slow, rapid, or pressure filters, using sand, anthracite, or other granular media singularly or in combination. Filter backwashing facilities remove trapped solids and restore the filter's effectiveness. *See* FILTRATION.

Disinfection. A primary drinking-water treatment objective is disinfection to destroy pathogens and other harmful organisms. Chlorination, using chlorine or chloramines, is the traditional disinfection technique, but chlorine dioxide is sometimes used. However, chlorine is not effective against some important pathogens. The use of chlorine in some waters produces chlorinated by-products, some of which are known or suspected carcinogens. Ozonation and ultraviolet (UV) irradiation are principal disinfection alternatives. Ozone is a strong oxidant and is effective against most pathogens. In the presence of bromide, ozonation of water can form bromate (BrO_3^-), which is a carcinogen. UV irradiation is very effective against most pathogens, including the chlorine-resistant pathogen *Cryptosporidium*. UV treatment is not known to produce any by-products of health significance.

Advanced processes. Other treatment processes have specialty applications for specific water quality challenges such as water hardness, iron and manganese that cause staining and color, contaminants that cause taste and odor, presence of natural compounds that form by-products during chlorination, and harmful organic or inorganic pollutants.

For treating high-mineral-content waters, ion exchange is most frequently used for softening, the process of removing calcium and magnesium of hard water. Ion exchange can also be used for deionization when ultrapure water is needed for industrial applications or for removing harmful inorganic pollutants such as arsenic. Reverse osmosis, electrodialysis, ultrafiltration, and other membrane processes are especially applicable for treating water that contains color and some complex organics or inorganics, as well as for desalination. *See* ION EXCHANGE; WATER SOFTENING.

Chemical precipitation is used in water softening, iron and manganese removal, and in other special cases for removing heavy metals. Enhanced coagulation can be used to remove natural organic matter and reduce the level of chlorinated organic contaminants formed during chlorination. Activated carbon (granular or powdered) is used for removing dissolved natural and synthetic organics, taste and odor compounds, and color. *See* ACTIVATED CARBON; PRECIPITATION (CHEMISTRY); WATER TREATMENT.

Oxidation processes, such as ozone, potassium permanganate, and chlorination, are used at the beginning of a water treatment plant to control taste and odor-causing compounds, reduce the microbiological load on the downstream treatment processes, and oxidize any reduced constituents. Advanced oxidation, using hydroxyl radicals produced from hydrogen peroxide in combination with either ozone or UV irradiation, may be implemented to destroy complex organic pollutants remaining in the water after filtration. Karl G. Linden; Robert A. Corbitt

Bibliography. American Society of Civil Engineers and American Water Works Association, *Water Treatment Plant Design*, 4th ed., 2004; American Water Works Association, *Water Quality and Treatment*, 5th ed., 1999; L. S. Clescerl, A. E. Greenberg, and A. D. Eaton (eds.), *Standard Methods for the Examination of Water and Wastewater*, 20th ed., 1999; R. A. Corbitt (ed.), *Standard Handbook of Environmental Engineering*, 2d ed., 1999; M. Watson Harza, *Water Treatment: Principles and Design*, 2d ed., 2005.

Water table

The upper surface of the zone of saturation in permeable rocks not confined by impermeable rocks. It may also be defined as the surface underground at which the water is at atmospheric pressure. Saturated rock may extend a little above this level, but the water in it is held up above the water table by capillarity and is under less than atmospheric pressure; therefore, it is the lower part of the capillary fringe and is not free to flow into a well by gravity. Below the water table, water is free to move under the influence of gravity. The position of the water table is shown by the level at which water stands in wells penetrating an unconfined water-bearing formation.

Where a well penetrates only impermeable material, there is no water table and the well is dry. But if the well passes through impermeable rock into water-bearing material whose hydrostatic head is higher than the level of the bottom of the impermeable rock, water will rise approximately to the level it would have assumed if the whole column of rock penetrated had been permeable. This is called artesian water, and the surface to which it rises is called the piezometric surface. *See* ARTESIAN SYSTEMS.

The water table is not a level surface but has irregularities that are commonly related to, though less pronounced than, those of the land surface. Also, it is not stationary but fluctuates with the seasons and from year to year. It generally declines during the summer months, when vegetation uses most of the water that falls as precipitation, and rises during the late winter and spring, when the demands of vegetation are low. The water table usually reaches its lowest point after the end of the growing season and its highest point just before the beginning of the growing season. Superimposed on the annual fluctuations are fluctuations of longer period which are controlled by climatic variations. The water table is also affected by withdrawals, as by

pumping from wells. *See* GROUND-WATER HYDROLOGY. Albert N. Sayre; Ray K. Linsley

Bibliography. W. Back (ed.), *Hydrogeology*, 1989; S. N. Davis and R. J. M. DeWiest, *Hydrogeology*, 1991; L. Huisman and T. N. Olsthorn, *Groundwater Recharge*, 1983; R. C. Pyne, *Groundwater Recharge Through Wells: A Guide to Aquifer Storage Recovery*, 1995.

Water treatment

Physical and chemical processes for making water suitable for human consumption and other purposes. Drinking water must be pathogenically safe, free from toxic or harmful chemical or substances, and comparatively free of turbidity, color, and taste- or odor-producing substances. Excessive hardness and high concentration of dissolved solids are also undesirable, particularly for boiler feedwater and industrial purposes. The treatment processes commonly used are sedimentation, coagulation, filtration, disinfection, softening, and aeration. A recently developed water treatment technology uses pressure-driven membranes as a barrier to pollutants.

Sedimentation. Silt, clay, and other fine material settle if the water is allowed to stand or flow at low velocity. Sedimentation occurs naturally in reservoirs and is accomplished in treatment plants using basins or settling tanks. The detention time in a settling basin may range from an hour to several days. The water may flow horizontally through the basin, with solids settling to the bottom, or may flow vertically upward at a low velocity so that the particles settle through the rising water. Settling basins are most effective if shallow and rarely exceed 3–6 m (10–20 ft) in depth. Basic sedimentation does not remove extremely fine or colloidal material within a reasonable time; therefore, it is used principally as a preliminary process to other treatment methods or following the coagulation–flocculation process.

Coagulation and flocculation. The chemical process (coagulation) of neutralizing the chemical and electrostatic forces that repel particles is followed by the physical process (flocculation) of encouraging particle collisions and agglomerating particles for successful removal by gravity sedimentation. Thus, extremely fine particles and colloidal material are combined into larger masses by this process. These masses, called floc, are large enough to settle in basins and to be caught on the surface of filters. Waters high in organic material and iron may coagulate naturally with gentle mixing. The term "coagulation" is usually applied to chemical coagulation, in which iron or aluminum salts are added to the water to form insoluble hydroxide floc. The floc is a feathery, absorbent substance to which color-producing colloids, bacteria, fine particles, and other substances become attached and are removed from the water.

The coagulant dose is a function of the physical and chemical character of the raw water, the adequacy of settling basins and filters, and the degree of purification required. Moderately turbid water coagulates more easily than perfectly clear water, but extremely turbid water requires more coagulant. Coagulation is more effective at higher temperatures. Lime, soda ash, or caustic soda may be required in addition to the coagulant to provide sufficient alkalinity for the formation of floc, and regulation of the pH (hydrogen-ion concentration) is usually desirable. Powdered limestone, clay, bentonite, or silica are sometimes added as coagulant aids to strengthen and weight the floc, and a wide variety of polymers developed in recent years are used for the same purpose.

Filtration. Suspended solids, colloidal material, bacteria, and other organisms are filtered out by passing the water through a bed of sand or pulverized coal, or through a matrix of fibrous material supported on a perforated core. Filtration of turbid or highly colored water usually follows sedimentation or coagulation and sedimentation. Soluble materials, such as salts and metals in ionic form, are not removed by filtration. *See* FILTRATION.

Slow sand filters. Used first in England around 1850, slow sand filters consist of beds of sand 51–122 cm (20–48 in.) deep, through which the water is passed at fairly low rates of $2.4–9.4 \times 10^4$ m^3/ha ($2.5–10 \times 10^6$ gal/acre). The bed sizes range from a fraction of an acre in small plants to several acres in large plants. An underdrain system of graded gravel and perforated pipes transmits the water from the filters to the point of discharge. The sand is usually fine, ranging 0.2–0.5 mm in diameter. The top of the filter clogs from bacterial growth with use, and a thin layer of dirty sand is scraped from the filter periodically to maintain proper flow-through capacity.

Slow sand filters operate satisfactorily with reasonably clear waters but clog rapidly with turbid waters. The filters are covered in cold climates to prevent the formation of ice and to facilitate operation in the winter. In milder climates, they are often open. Slow sand filters have a high bacteriological efficiency, but few have been built since the development of water disinfection, because of the large area required, the high construction cost, and the labor needed to clean the filters and to handle the filter sand. Slow sand filters are still used in many English and European cities, but have not been built in the United States since 1950, and few remain in operation.

Rapid sand filters. These operate at rates of $1.17–2.34 \times 10^6$ m^3/(ha/day) [$1.25–2.5 \times 10^8$ gal/(acre/day)], or 25 to 50 times faster than the slow sand filters. The high rate of operation is made possible by coagulation and sedimentation ahead of filtration to remove the heaviest part of the load, the use of coarse sand, and facilities for backwashing the filter to keep the bed clean. The filter beds are small, generally ranging from 14 m^2 (150 ft^2) in small plants to 140 m^2 (1500 ft^2) in the largest plants. The filters consist of a layer of sand or, occasionally, crushed anthracite coal 46–62 cm (18–24 in.) deep, resting on graded layers of gravel above an underdrain system. The sand is coarse, 0.4–1.0 mm in diameter, depending upon the raw water quality and pretreatment, but the grain size must be fairly uniform to assure proper

backwashing. The underdrain system serves both to collect the filtered water and to distribute the wash water under the filters when they are being washed. Several types of underdrains are used, including perforated pipes, perforated false bottoms of concrete, and tile and porous plates.

Filters are backwashed at rates 5–10 times the filtering rate. The wash water passes upward through the sand and out of the filters by way of wash-water gutters and drains. Washing agitates the sand bed and releases the dirt to flow out of the filter with the wash water. The quantity of water used for washing ranges from 1 to 10% of the total output, depending upon the turbidity of the water applied to filters and the efficiency of the filter design. Combination air and water filter washes are popular in Europe but are not often used in the United States.

Municipal and large-capacity filters for industry usually are built in concrete boxes or in open tanks of wood and steel. The flow through the sand may be caused by gravity, or the water may be forced through the sand under pressure by pumping. Pressure filters can be operated at higher rates than gravity filters, because of the greater head available to force the water through the sand. However, excessive pressure may increase the effluent turbidity, and bacteria may appear in the discharge water. For this reason, and because pressure filters are difficult to inspect and keep in good order, open gravity filters are favored for public water supplies.

Diatomaceous earth filters. Swimming-pool installations and small water supplies frequently use this type of filter. The filters consist of a medium or septum supporting a layer of diatomaceous earth through which the water is passed. A filter layer is built up by the addition of diatomaceous earth to the water. When the pressure loss becomes excessive, filters must be backwashed and a fresh layer of diatomaceous earth applied. Filter rates of $1.7–4.1 \times 10^{-3}$ $m^3/(s/m^2)$ [2.5–6 gal/(min/ft^2)] are attained.

Disinfection. There are several water treatment methods designed to kill harmful microorganisms, particularly pathogenic bacteria, spores, viruses, and protozoa. The application of chlorine or chlorine compounds is the most common method; however, its use is under increasing scrutiny. Emerging disinfection methods include the use of ultraviolet light and ozone, while boiling is used as a household emergency measure.

Chlorination is practiced almost universally in public water supplies in North America. It is sometimes the sole treatment of clear, uncontaminated waters. In conventional water treatment plants, it is used along with coagulation–flocculation and filtration. Chlorination, in the form of free chlorine or chloramines, is used also to protect against contamination of water in distribution mains and reservoirs after purification.

Chlorine gas is relatively economical and easy to apply in large systems. For small works, calcium hypochlorite or sodium hypochlorite is frequently used. Regardless of which form is used, the dose varies with the water quality and degree of contamination. Clear, uncontaminated water can be disinfected with small doses, usually less than 1 part per million, whereas contaminated water may require several times as much. The amount of chlorine taken up by organic matter and minerals in water is known as the chlorine demand. For proper disinfection, the dose must exceed the demand so that free chlorine remains in the water for adequate contact time to achieve the desired disinfection. Chlorine is applied before filtration, after filtration, or at both times. *See* CHLORINE; HYPOHALOUS ACID.

Chlorination alone is not reliable for the treatment of contaminated or turbid water. A sudden increase in the chlorine demand may absorb the full dose and provide no residual chlorine for disinfection, and it cannot be assumed that the chlorine will penetrate particles of organic matter, which may shield pathogens.

Although chlorine is currently an integral part of most water treatment plants, there are a number of concerns related to the use of chlorine. Chlorine sometimes causes objectionable tastes or odors in water. This may be due to an excessive chlorine dose, but more frequently it is caused by a combination of chlorine and organic matter, such as algae, in the water. Some algae, relatively unobjectionable in the natural state, produce unbearable tastes after chlorination. In other cases, strong chlorine doses oxidize the organic matter completely and produce odor-free water. When chlorine combines with natural organic matter present in the water, a number of chlorinated by-products can be formed, many of which are known or suspected carcinogens. Some of these disinfection by-products are regulated, and water plants need to meet strict standards to adhere to public health goals. In addition to formation of by-products, chlorine is not effective against some important protozoan pathogens. For these reasons, disinfection alternatives such as ultraviolet (UV) irradiation or ozone are supplementing chlorination. Excessive chlorine may be removed by dechlorination with sulfur dioxide. However, for protection of the distribution system, it is important to maintain some residual chlorine all the way to the consumer's tap. To aid in maintaining residual chlorine, ammonia is often added to create chloramines, which are more stable over time than free chlorine. The formation of chloramines has the added advantage of minimizing formation of by-products in the distribution system.

Ozonation and UV irradiation are the principal disinfection alternatives to chlorination. Ozone is a strong oxidant produced in a gaseous form and bubbled into water, where it dissolves. Ozone is highly reactive and is consumed quickly during treatment. The mechanism of ozone attack on pathogens is oxidation of the outer membrane or of the nucleic acids. Ozone is very effective against most pathogens; however, cold temperature can negatively affect its performance for some pathogens, such as *Cryptosporidium* oocysts. In the presence of bromide, ozonation of water can form bromate (BrO_3^-), which is a carcinogen. Bromate formation can be minimized through appropriate engineering design. UV

irradiation is produced from mercury vapor lamps, similar to household fluorescent lamps. These lamps are suspended in the water that flows through the system. A UV contact time of only a few seconds is needed for water disinfection. UV light emitted from the lamps in the wavelength range of 240–280 nanometers is very effective against most pathogens, with peak effectiveness around 260 nm, which coincides with the peak absorbance of deoxyribonucleic acid (DNA). The mechanism of action for UV disinfection is the transfer of UV energy from the photon to the microorganism's DNA, damaging sites that render the pathogen unable to reproduce, and thus effectively dead. UV has become of great interest to the water treatment industry for disinfection because it is extremely effective against the chlorine-resistant pathogen *Cryptosporidium*. UV treatment is not known to produce any by-products of health significance. One drawback with both ozone and UV energy is that they do not stay in the water to protect it as it moves from the treatment plant through the distribution system. *See* OZONE; ULTRAVIOLET RADIATION.

One of the functions of environmental engineers in protecting public health through water treatment is to balance the chemical risks from using too much chlorine with the microbial risks of not using enough. This balance of risks can be mitigated through the tailored addition of other disinfection processes, such as UV irradiation and ozone, which complement the effectiveness of chlorine. A multiple-barrier disinfection approach will likely be used more frequently in the future to help minimize by-products and optimize killing of all types of waterborne pathogens.

Granular activated carbon. Fine-grain activated carbon has been used for years to reduce objectionable tastes and odors, harmful organic compounds, and heavy metals in public water supplies. Since the capacity of granular activated carbon (GAC) for this use is great, no special facilities are normally designed for taste and odor control. Instead, the carbon is put on top of the filters or is used as the filter medium during rapid filtration.

Granulated activated carbon has been proposed by the Environmental Protection Agency (EPA) to eliminate or reduce potentially carcinogenic organic substances in public water supplies. The harmful substances, such as chlorinated organic by-products, are derived from naturally occurring humic and fulvic acids combining with chlorine and bromine. A nationwide survey of chlorinated public water supplies indicated many contained regulated and unregulated chlorinated by-products. As a step toward correcting the situation, the EPA has developed federal regulations limiting the maximum concentration of two classes of chlorinated by-products, trihalomethanes and haloacetic acids, to 0.08 and 0.06 mg per liter, respectively, in the distribution system. Communities not meeting these limitations must adjust their disinfection procedures in a manner that decreases the by-products to acceptable levels. Where such adjustments are found to be insufficient, granulated activated carbon facilities or disinfection alternatives are being installed. GAC facilities have been built at only a few places, and there is little reliable operating and cost data. *See* HALOGENATED HYDROCARBON; HUMUS.

Designing the contact chamber to pass water plant effluent through a bed of granulated activated carbon is relatively simple. The real job and cost of the process are in the periodic removal, cleaning, regeneration, and returning of the granulated activated carbon to the contact chambers, and adding new material to make up for losses. Several regeneration furnaces are available, including infrared furnace, stream, and multiple hearths. All of these involve the operation and maintenance of machinery and equipment, and a competent labor force. The frequency of regeneration varies greatly, since it depends not only upon how much organics must be removed, but upon what other natural or synthetic chemicals are present in the water that interfere with organics removal, and the type of granulated activated carbon used.

With these factors in mind, the overall cost of facilities and operations can be as high as the cost of conventional water treatment. Since changes in raw water quality may affect the performance of the granulated activated carbon, some irregularity can be expected. Design, construction, and installation of equipment and monitoring facilities should not be started without reliable, expert advice and some bench or pilot testing. *See* ACTIVATED CARBON.

Water softening. Water softening is the process of removing the "hardness" caused by calcium and magnesium salts. These salts make washing difficult, waste soap, and cause unpleasant scums and stains in households and laundries. They are especially harmful in boiler feedwater because of their tendency to form scales.

Municipal water softening is common where the natural water has a hardness in excess of 150 parts per million. Two methods are used: (1) The water is treated with lime and soda ash to precipitate the calcium and magnesium as carbonate and hydroxide, after which the water is filtered. (2) The water is passed through a porous cation exchanger which has the ability of substituting sodium ions in the exchange medium for calcium and magnesium in the water. The exchange medium may be a natural sand known as zeolite or may be manufactured from organic resins. It must be recharged periodically by backwashing with brine.

For high-pressure steam boilers or some other industrial processes, almost complete deionization of water is needed, and treatment includes both cation and anion exchangers. Lime-soda plants are similar to water purification plants, with coagulation, settling, and filtration. Zeolite or cation-exchange plants are usually built of steel tanks equipped for backwashing the media with salt brine. If the water is turbid, filtration ahead of zeolite softening may be required. *See* ION EXCHANGE; WATER SOFTENING; ZEOLITE.

Aeration. Aeration is a process of exposing water to air by dividing the water into small drops, by

forcing air through the water, or by a combination of both. The first method uses jets, fountains, waterfalls, and riffles; in the second, compressed air is admitted to the bottom of a tank through perforated pipes or porous plates; in the third, drops of water are met by a stream of air produced by a fan.

Aeration is used to add oxygen to water and to remove carbon dioxide, hydrogen sulfide, and taste-producing gases or vapors. Aeration is also used in iron-removal plants to oxidize the iron ahead of the sedimentation or filtration processes. *See* WATER POLLUTION.

Karl G. Linden; Richard Hazen

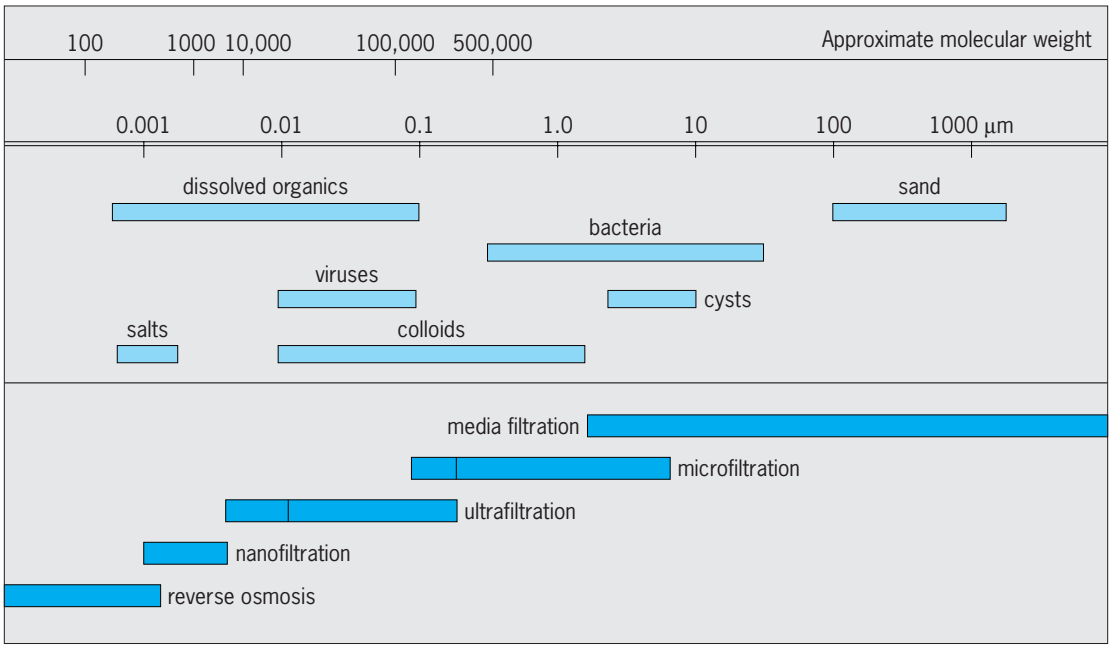
Pressure-driven membranes. Membranes act as selective barriers to pollutants, where some constituents pass through while others are blocked. To move materials across a membrane, a driving force, such as pressure, is needed. Pressure-driven membranes have recently been developed and implemented as a water treatment technology for separating suspended and dissolved materials from pure water. Pressure-driven membranes are categorized as low-pressure and high-pressure membranes. Low-pressure membranes include microfiltration and ultrafiltration. High-pressure membranes include nanofiltration and reverse osmosis.

Membranes can operate under positive or negative (vacuum) pressure and are typically characterized by their molecular weight cut-off (MWCO), which is the molecular mass of the chemical for which the membrane has a rejection capability of greater than 90%. They are sometime also characterized by pore size, referring to the diameter of the micropores in a membrane surface material, or the size of the particulate that can be retained. The **illustration** presents a number of different membranes and the types of pollutants that can be removed. The material that a membrane is made of can also affect removal of a contaminant due to chemical interactions.

Membranes are configured in modules, consisting of individual membrane fibers bundled together in a cartridge. A membrane treatment system has a number of modules arranged on a rack, sized by the number of modules per flow to be treated. In addition to the modules, membrane systems need to have provisions for backwashing and chemical cleaning. The membrane fibers may treat water in a cross-flow manner where water flows tangentially to the membrane surface, or a dead-end configuration where all water passes through the membrane. When the pressure across the membrane builds up, the membranes are backwashed and returned to the starting pressure. This backwashing may occur every 20 to 60 min for low-pressure membranes. Chemical cleaning is necessary when the pressure does not return to baseline values over time.

Microfiltration. Membranes that are used mainly for separation of particles, larger microorganisms, and minor removal of dissolved organic matter or pollutants are microfiltration (MF) membranes. These can replace or augment traditional sand media filtration. Depending upon the chemical characteristics of the membrane material, some dissolved matter may be removed. The pore size of microfiltration membranes is typically around 1 micrometer, which will remove larger pathogens.

Ultrafiltration. A low-pressure membrane with a smaller pore size is the ultrafiltration (UF) membrane. This membrane is also typically used for removing suspended and colloidal particles and microorganisms, but is more efficient at removing smaller materials with a size around 0.05 μm. Thus, ultrafiltration can remove bacteria and some larger viruses, as well as some dissolved matter. Sometimes ultrafiltration is used in combination with microfiltration, which provides preremoval of larger particles, minimizing



Molecular weight cut-off and pore sizes for membranes relative to various water quality concerns. (*EPA 815-C-01-001*)

clogging and frequency of backwashing and cleaning for the ultrafiltration membrane.

Nanofiltration. A high-pressure membrane that provides very good removal of many dissolved contaminants and viruses is the nanofiltration membranes. Nanofiltration (NF) removes molecules in the nanometer range and with a molecular weight greater than 1000 daltons. Thus, NF is capable of removing some elements associated with hardness (Ca and Mg), as well as bacteria and most viruses. Nanofiltration can also remove pesticides and other organic contaminants from water, as well as natural organics that can lead to the formation of chlorinated organic by-products. The charge of the particles can affect the removal efficiency, and small-molecule pollutants are typically not removed well.

Advanced treatment. Some treatment processes have specialty applications for removing or destroying contaminants that cause an undesirable taste and odor, natural compounds that may lead to formation of by-products during chlorination, and harmful organic or inorganic pollutants. Although not a threat to human health, taste- and odor-causing compounds, such as geosmin (musty odor) and methylisoborneol (MIB), derived from algae in water can lead to major esthetic problems with water and the perception that it is not fit to drink. If not removed, natural organic matter in surface waters can serve as precursor to the formation of chlorinated by-products. Organic contaminants, such as synthetically derived pesticides, pharmaceutical and personal care products, endocrine-disrupting contaminants, and industrial chemicals and solvents, are known or suspected to be harmful to human and aquatic health and are sometimes found in raw drinking water supplies. Inorganic contaminants, such as arsenic, perchlorate, nitrate, and chromium, can enter the water naturally or from a waste source. Because many of these contaminants can be difficult to remove using typical water treatment processes, advanced treatment is needed where their presence is a water-quality or health threat.

Ion exchange, for treating high-mineral-content waters, is most frequently used for softening. Ion exchange can also be used for deionization when ultrapure water is needed for industrial applications or removal of harmful inorganic pollutants such as arsenic. Reverse osmosis, electrodialysis, nanofiltration, and other membrane processes are especially applicable for treatment of water sources containing color and some complex organics or inorganics, including desalination. *See* MEMBRANE SEPARATIONS; WATER DESALINATION.

Chemical precipitation is used in water softening, iron and manganese removal, and other special cases for removal of heavy metals. Enhanced coagulation can be used to remove natural organic matter to reduce the level of chlorinated organic contaminants formed. Activated carbon (granular or powdered) applications are used for removal of dissolved natural and synthetic organics, taste and odor compounds, and color. *See* PRECIPITATION (CHEMISTRY).

Oxidation processes, such as ozone, potassium permanganate, and chlorination, are used at the beginning of a water treatment plant to control taste- and odor-causing compounds, reduce the microbiological load on the downstream treatment processes, and oxidize any reduced constituents. Ozone or advanced oxidation using hydroxyl radicals, produced from hydrogen peroxide in combination with either ozone or UV irradiation, may be implemented to destroy the numerous complex organic pollutants remaining in the water after filtration. *See* ENVIRONMENTAL ENGINEERING; WATER SUPPLY ENGINEERING.

Karl G. Linden

Bibliography. American Society of Civil Engineers and American Water Works Association, *Water Treatment Plant Design*, 4th ed., 2004; American Water Works Association, *Water Quality and Treatment*, 5th ed., 1999; L. S. Clescerl, A. E. Greenberg, and A. D. Eaton (eds.), *Standard Methods for the Examination of Water and Wastewater*, 20th ed., 1999; R. A. Corbitt (ed.), *Standard Handbook of Environmental Engineering*, 2d ed., 1999; M. Watson Harza, *Water Treatment: Principles and Design*, 2d ed., 2005.

Water-tube boiler

A steam boiler in which water circulates within tubes and heat is applied from outside the tubes. The outstanding feature of the water-tube boiler is the use of small tubes exposed to the products of combustion and connected to steam and water drums which are shielded from these high-temperature gases. Thus, possible failure of boiler parts exposed to direct heat transfer is restricted to the small-diameter tubes and, in the event of failure, the energy released is reduced and explosion hazards are minimized.

The water-tube construction facilitates greater boiler capacity by increasing the length and the number of tubes and by using higher pressure, since the

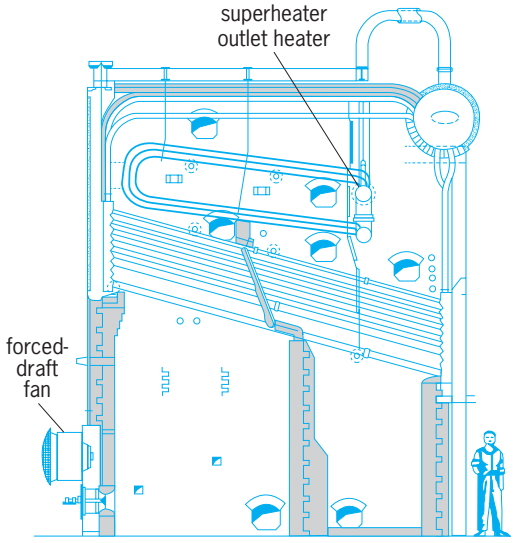


Fig. 1. Straight-tube-type boiler.

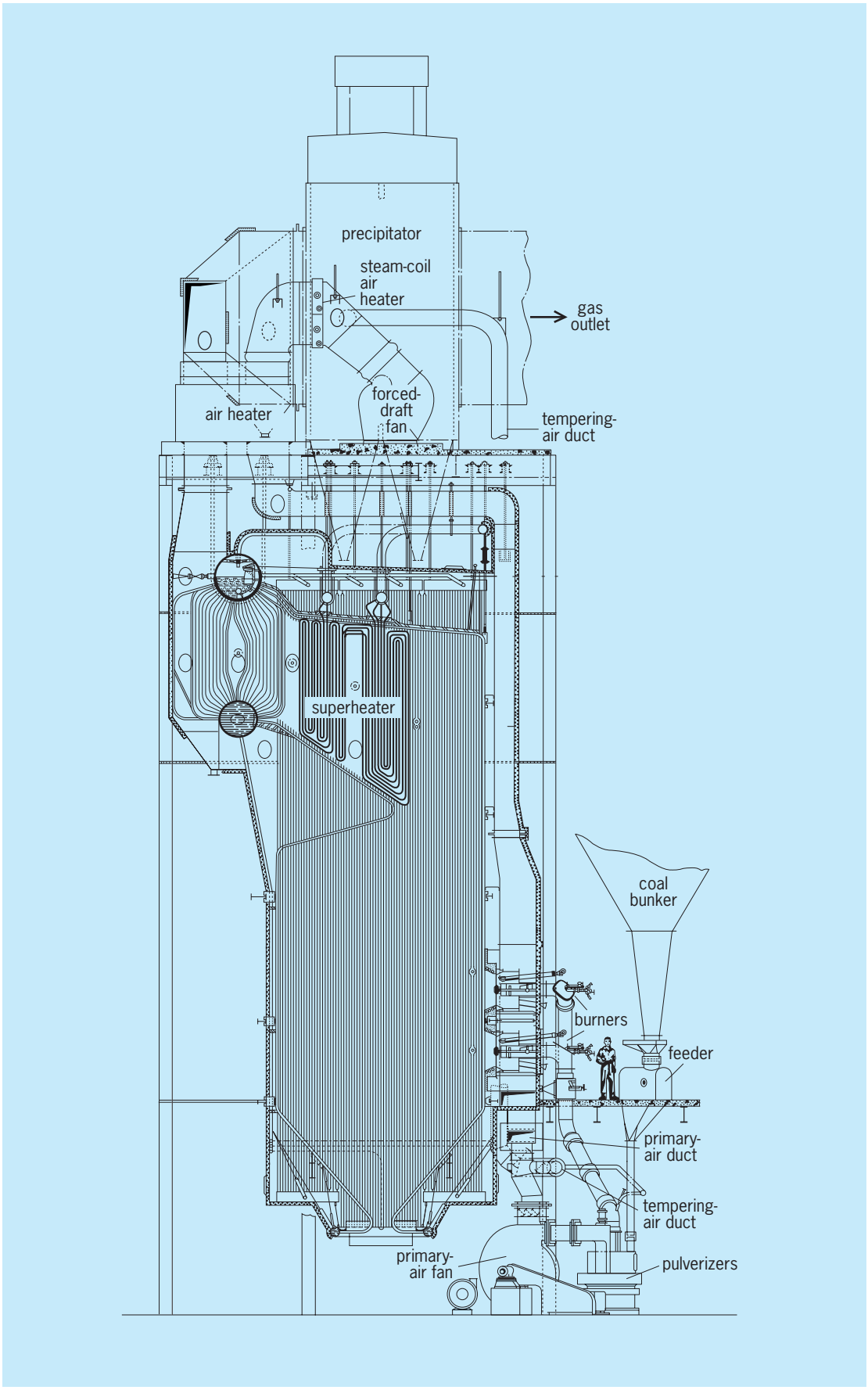


Fig. 2. Bent-tube-type boiler.

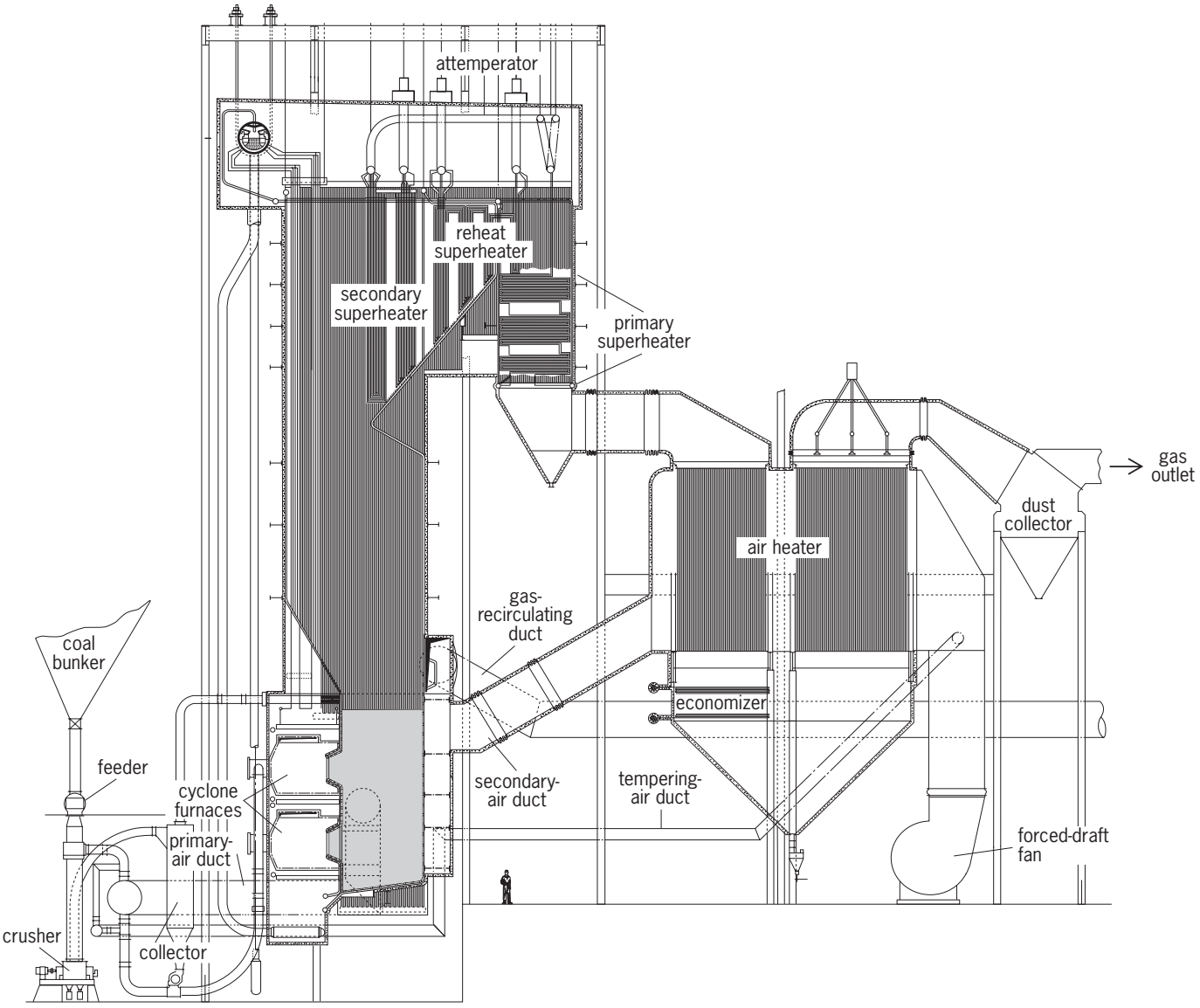


Fig. 3. Radiant-type boiler.

relatively small-diameter tubes do not require an abnormal increase in thickness as the internal pressure is increased. In addition, water-tube boilers offer great versatility in arrangement, and this permits efficient use of the furnace, superheater, and other heat-recovery components.

There are many types of water-tube boilers but, in general, they can be grouped into two categories: the straight-tube and the bent-tube types. In essence, both types consist of banks of parallel tubes which are connected to, or by, headers or drums. The early water-tube boilers were fitted with refractory furnaces. However, most modern water-tube boilers utilize a water-cooled surface in the furnace, and this surface is an integral part of the boiler's circulatory system. Further, modern water-tube boilers generally incorporate the use of superheaters, economizers, or air heaters to utilize more efficiently the heat from the fuel and to provide steam at a high potential for useful work in an engine or turbine. *See* BOILER ECONOMIZER; SUPERHEATER.

Straight-tube boiler. The straight-tube boiler (**Fig. 1**), often called the header-type boiler, has the advantage of direct accessibility for internal inspection and cleaning through handholes, located opposite each tube end, in the headers. The steam-generating sections are joined to one or more steam-and-water drums located above and parallel or transverse to the boiler tube bank. The circulation of water in the downcomer headers and of the water-steam mixture in the boiler tubes and riser headers is the result of the differential density between the water in the downcomers and the steam and water mixture in the heated boiler tubes and riser portion of the circuit. In many designs the tube bank can be baffled to increase the rate of flow of the products of combustion and, thus, improve heat transfer and the resultant absorption efficiency.

The straight-tube boiler is not applicable to high-pressure designs because of header limitations, and capacity is restricted by space requirements. Thus,

bent-tube instead of header-type boilers are used in most modern industrial installations.

Bent-tube boiler. In bent-tube boilers (**Fig. 2**), commonly referred to as drum-type boilers, the boiler tubes terminate in upper and lower steam and water drums which have few access openings. Although internal inspection is restricted, the tubes can be mechanically or chemically cleaned, and developments in water treatment and cleaning methods have overcome the early objections to the use of bent tubes. Circulation in drum-type boilers, as in header-type boilers, is due to the difference between the density of the water in the downcomer tubes and the water-steam mixtures in the riser tubes.

Because of the greater slope of tubes, or even the use of vertical tubes, in drum-type boilers, there is less possibility of steam pockets forming in the tubes and, thus, the rate of heat absorption can be increased appreciably. Consequently, drum-type boilers can be used for both low and high pressures and for capacities ranging from a few thousand to more than a million pounds of steam per hour. In addition, they can be designed to meet almost any space allotment and can accommodate heat-absorbing components such as superheaters, reheaters, economizers, and air heaters. An evolution of the bent-tube boiler is the radiant-type boiler in which the boiler tubes (steam-generating surface) form the boundary of the furnace and the containment for the superheater, reheater, and economizer surface (**Fig. 3**). Such boilers can deliver several million pounds of steam per hour at high steam pressures and temperatures. Further, by eliminating the steam drum and using a once-through flow of water and steam, they can operate at supercritical steam pressures. *See* BOILER; BOILER FEEDWATER; STEAM; STEAM-GENERATING FURNACE; STEAM-GENERATING UNIT.

George W. Kessler

Bibliography. E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; A. G. Blokh, *Heat Transfer in Steam Boiler Furnaces*, 1987; J. J. Jackson, *Steam Boiler Operation: Principles and Practices*, 2d ed., 1987.

Water tunnel (research and testing)

A hydrodynamic facility used for research, test, and evaluation, comprising a well-guided and controlled stream of water in which items for test are placed. The water tunnel is in many ways similar in appearance, arrangement, and operation to a subsonic wind tunnel. It is related to and complementary to the towing tank, in which the test item, usually a scale model of a ship or ship component, is towed through stationary water and evaluated through observation and measurement. In a water tunnel the test item is held stationary while the water is circulated around it. Many water tunnels are capable of operation with variable internal pressure to simulate the phenomenon of cavitation. *See* CAVITATION; TOWING TANK; WIND TUNNEL.

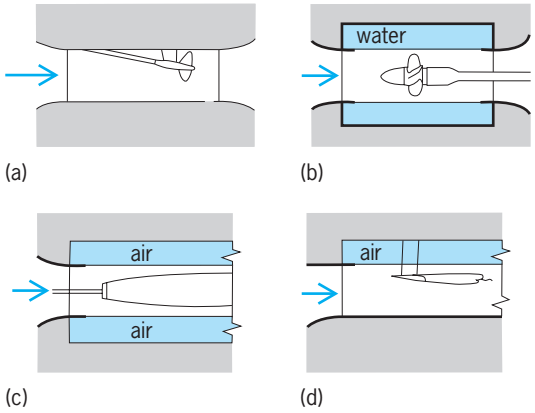


Fig. 1. Types of water-tunnel test sections with typical models. (*a*) Closed throat. (*b*) Open throat. (*c*) Free jet. (*d*) Free surface.

Classification. Water tunnels may be classified, in part, by the type of test section used. The most common section is the closed throat (**Fig. 1*a***) in which the test section flow has solid boundaries. The advantage of this arrangement is its simplicity and efficiency, but the model must be small relative to the tunnel cross section to avoid large wall effects. Small wall effects are theoretically correctable. In an open-throat test section (Fig. 1*b*) the water jet passes through a water-filled chamber of larger diameter. This minimizes wall effects, and many tunnels dedicated to propeller testing use such an arrangement. When very low test section cavitation numbers are required, or for fully cavitating flows, a free jet (Fig. 1*c*) in which the water jet passes through an air-filled chamber is useful. However, capture of the free jet and removal of excess entrained air prior to recirculation is not easily achieved over a broad range of test conditions. To study cavity flows on surface-piercing components or hydrofoils which operate near a free surface, a free-surface tunnel (Fig. 1*d*) is required in which three sides of the water flow are bounded by solid walls and the upper surface is open to air at controlled pressure levels. This arrangement is often referred to as a water channel. *See* HYDROFOIL CRAFT; OPEN CHANNEL.

Another distinction among water tunnel types is whether or not they recirculate the flow. If the tunnel is nonrecirculating, water may "blow down" from a pressurized or elevated water tank or water may be diverted from a continuous source such as a waterfall or dam on a river. The blow-down tunnel has a limited test period proportional to the size of the storage tank. All water tunnels make use of transparent test-section viewing windows.

Construction. The largest water tunnel in the world is the William B. Morgan Large Cavitation Channel, which is operated by the U.S. Navy and is shown in **Fig. 2**. Typical of most designs, it is of the recirculating, closed-throat type with a carefully designed contraction ahead of the test section. An axial-flow pump located in the lower horizontal leg impels water through the circuit and test section. The strong contraction ensures good velocity uniformity in the

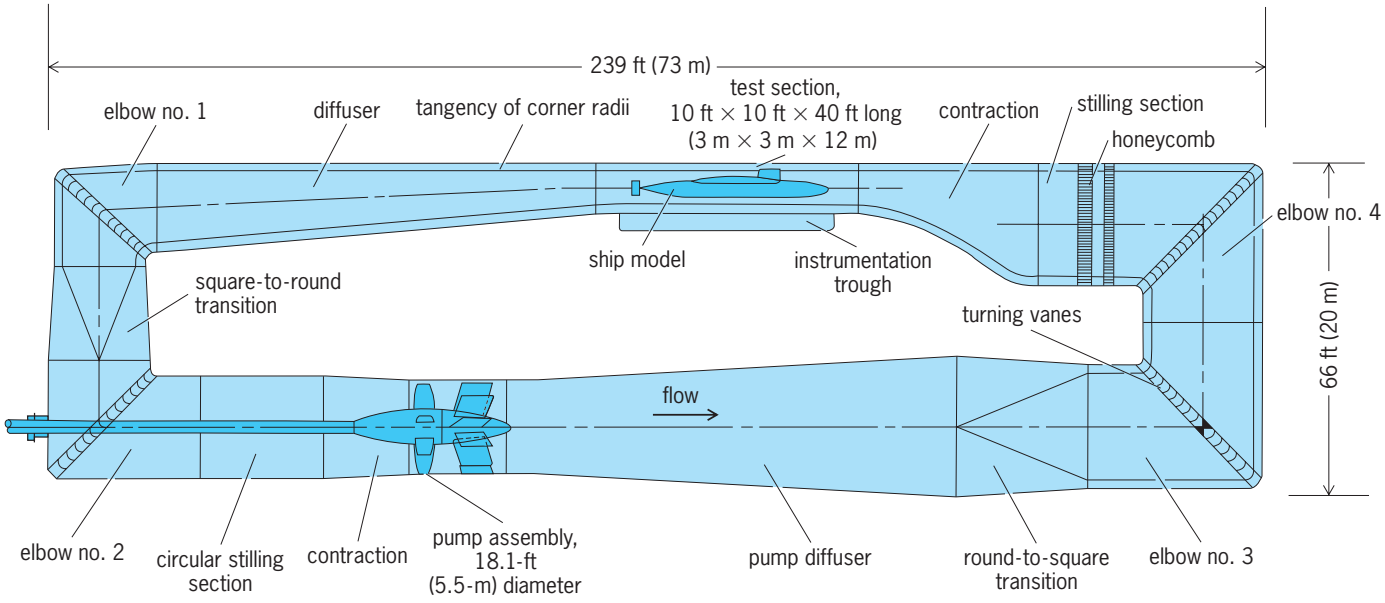


Fig. 2. Elevation of the William B. Morgan Large Cavitation Channel operated by the U.S. Naval Surface Warfare Center, Carderock Division. The water tunnel is located in Memphis, Tennessee.

test section, by the change of potential energy (pressure), which is uniform across the contraction entrance, to kinetic energy (velocity) of the test section. Turbulence is reduced by passage through a stilling section containing two honeycombs. The size of a water tunnel is generally specified in terms of the dimensions of the test section where measurements are performed. The tunnel shown in Fig. 2 has a rectangular test section of 10 ft by 10 ft by 40 ft long (3 m by 3 m by 12 m). *See* FLUID-FLOW PRINCIPLES; TURBULENT FLOW.

After passing through the text section, the velocity of the water in a recirculating tunnel is reduced to avoid cavitation and excess energy losses in other parts of the loop. Reduction of velocity takes place in the diffuser, along with an increase in pressure which helps to suppress possible cavitation of the turning vanes. Turning vanes serve to direct the flow uniformly and with minimal losses through a 90° turn at each elbow. Relatively uniform flow input to the pump is required to decrease the likelihood of undesirable vibration and pump blade cavitation.

Pressure control. To observe the cavitation characteristics of propellers and other test bodies, the water tunnel internal pressure must be varied independently of the velocity in the test section. Generally, for a model of a surface ship tested at speeds lower than full scale, it is necessary to decrease pressure to match full-scale cavitation conditions. If the model test speed is higher than full scale, the pressure must be increased. For submarines and similar applications which experience high full-scale operating pressures, it is also necessary to increase the tunnel pressure above the normal atmospheric level.

Closed- and open-throat tunnels. Most tunnels are of the closed- or open-throat type, operated with the entire test circuit filled with water. Consequently, they are well suited to investigations of completely submerged bodies or of conduits and turbomachinery such as pumps or turbines that are completely filled with water. Tunnels dedicated to turbomachinery are frequently referred to as test stands. Surface-ship hulls can be tested mounted flush to the top of a water tunnel test section, with the normally submerged portion of the hull exposed to the flow. The rigid wall or a specially constructed ground board takes the place of the free surface.

Free-surface tunnels. Often the behavior of the water surface around the ship has an important effect on the cavitation characteristics of the propeller or other appendages under test. For these cases, tests must be performed in the free-surface type of tunnel, or channel. The principal problem associated with free-surface tunnels is elimination of excess entrained air from the recaptured water downstream of the test section. For this reason, several free-surface tunnels have been designed to discharge water directly into a large reservoir after passage through the test section. A common practice for recirculating free-surface tunnels is to skim the top 10% of the flow from the test section and divert this to a reservoir, while recirculating the remainder after passage through a deaeration section.

Air content control. In closed or open types of tunnels, control of air content in the form of bubbles and dissolved gases is considered to be critical for the accurate replication of cavitation phenomena at model scale. Total gas content may be reduced by exposing the circulating water to reduced ambient pressure for an extended period of time. Dissolved air gradually comes out of solution to form bubbles which can be drawn off. Residual bubbles, or bubbles generated by the test body, can be removed during circulation in the tunnel by two processes: resorption and collection. Resorption takes place when part of the circuit is at a much lower height (higher static pressure) than the test section. The increased pressure forces free bubbles into solution. An alternative method is to slow the flow to a very low speed just upstream or downstream of the test section, and to collect free

bubbles in trays or tubes and guide them to air evacuation points. Total removal of dissolved gas and bubbles is not desired for cavitation tests. Rather, a stable bubble population which allows good correlation of cavitation inception and cavitation patterns with realistic prototype conditions is sought. Although not a feature of most existing water tunnels, aeration devices to inject desired bubble populations are being incorporated into the design of some tunnels. Because bubbles are elastic systems, they can absorb acoustic energy. For measurements of noise generated by noncavitating bodies, it is especially important to control that part of the free bubble population which is responsible for absorption of radiated noise in the frequency range of interest.

Applications. The water tunnel is applied in much the same way as a subsonic wind tunnel, the principal difference being the possibility of vaporization of the water due to local decrease in pressure (cavitation) or the existence of a free surface and attendant interface phenomena such as waves, ventilation, and spray. Water tunnels are used to investigate the dynamics, hydrodynamics, and cavitation of submerged and semisubmerged bodies such as propellers, ships, submarines, torpedoes, and hydrofoils and of turbomachinery. They are also indispensable to research on general flow phenomena in liquids. An application of great importance has been the acoustic characterization of propellers under both cavitating and noncavitating conditions. This requires low background-noise level in the frequency range of interest. Thus, turning vanes, diffuser, and impeller must be designed to strict acoustic as well as hydrodynamic criteria. *See* HYDRODYNAMICS.

The first use of water tunnels was to identify the onset of thrust breakdown of propellers due to excessive cavitation. This is still an important application, extending to other types of marine propulsors and turbomachinery as well. In a typical test of a propeller, the thrust and torque are measured as a function of rate of rotation and speed of advance through the water, for various pressures. Cavitation inception is observed, visually using stroboscopic lighting and acoustically using hydrophones. When this information is combined with that derived from towing-tank tests of the hull alone and in combination with a propeller, the thrust, torque, and power characteristics of the full-scale ship, including the operating point of best efficiency, can then be predicted. *See* HYDROPHONE; PROPELLER (MARINE CRAFT); STROBOSCOPIC PHOTOGRAPHY.

In sufficiently large water tunnels, such as the one in Fig. 2, the propeller can be tested with a completely appended model hull. This provides a more realistic operating environment. It also enables pressures induced on the hull by the propeller to be measured, and full-scale vibration levels to be predicted. The acoustic radiation of the propeller is also influenced by the presence of the hull. Noise and vibration characteristics of the hull-propeller combination are important for military surface ships and submarines because of the widespread use of acoustic detection techniques. For merchant ships, noise and vibration are important for habitability and avoidance of maintenance of vibration-sensitive components. *See* ACOUSTIC NOISE; MECHANICAL VIBRATION.

Limitations. The limitations of water tunnels in some respects are similar to those of wind tunnels. To scale dynamic effects (such as the flow induced forces) between the model and full-scale test objects, it is desirable to match the Reynolds numbers, Re, between the model and full-scale flows. The Reynolds number is defined as $\mathrm{Re} = UL\rho/\mu$, where U is a reference flow speed (typically the average upstream flow speed), L is a reference length (typically based on the size of the object being considered), and ρ and μ are the liquid density and dynamic viscosity, respectively. For flows with liquid-gas interfaces, the Froude number, Fr, which is defined as $\mathrm{Fr} = U/(gL)^{1/2}$, may also be considered. Here, g is the gravitational acceleration. In general, it is not possible to match both Re and Fr of test and full-scale flows, but is it often sufficient to conduct the model test above a threshold value of the Reynolds number while attempting to match the Froude number. *See* FROUDE NUMBER; REYNOLDS NUMBER.

Since the model scale is usually smaller in size than the full scale, the model-scale flow velocity might be increased to achieve dynamic similarity. Besides the obvious increase in power that increased velocity requires, it becomes necessary to increase pressure to maintain cavitation similitude. The cavitation number, σ, is defined as $\sigma = (P - P_v)/(1/2)\rho U^2$, where P is a reference pressure (typically the static pressure of the flow upstream of the test object), and P_v is the vapor pressure of the liquid. To achieve cavitation similarity, it is desirable to match the cavitation number between the model and full-scale flows. The need to increase the pressure at the model scale increases the required strength (shell thickness) of the water tunnel structure. This pressure constraint may be more restrictive than the power constraint, as pressure must rise with the square of the velocity. Therefore, water channel speeds are generally restricted to less than 50 knots (25 m/s) for relatively large tunnels. *See* DIMENSIONLESS GROUPS; DYNAMIC SIMILARITY; MODEL THEORY. Richard Stone Rothblum; Robert J. Etter; Steven L. Ceccio

Bibliography. C. E. Brennen, *Cavitation and Bubble Dynamics*, Oxford University Press, 1995; W. F. Brownell, *Two New Hydromechanics Research Facilities at the David Taylor Model Basin*, DTMB Rep. 1690, December 1962; R. J. Etter and M. B. Wilson, The large cavitation channel, in *Proceedings of the 23rd American Towing Tank Conference*, New Orleans, 1992; R. J. Etter et al., High Reynold's number experimentation in the U.S. Navy's William B. Morgan Large Cavitation Channel, *Meas. Sci. Technol.*, 16:1701–1709, 2005; H. D. Harper, A modernized control system for the DTNSRDC 36 inch variable pressure water tunnel, *Proceedings of the 20th American Towing Tank Conference*, Hoboken, 1983; E. V. Lewis (ed.), *Principles of Naval Architecture*, 3 vols., 2d ed., Society of Naval Architects and Marine Engineers, 1990; H. Lindgren and E. Bjärne, *Ten Years of Research in the SSPA Large Cavitation Tunnel*, SSPA Publ. 86, 1980.

Watermelon

The edible fruit of *Citrullus lanatus*, of the family Cucurbitaceae. The plant is an annual prostrate vine with multiple stems alternating from short nodes near the base of the main axis, reaching lengths of 10 to 15 ft (3 to 4.5 m). Short vine or "bush" type cultivars exist but are not commercially useful. In a typical monoecious plant, pistillate flowers occur at every seventh node, with staminate flowers at intervening nodes. Cultivated plants range from 250 to 1000 hills per acre (625 to 2500 hills per hectare) according to variety and region. Commercially, an average yield of one mature melon per plant is acceptable.

Varieties and production. The numerous open-pollinated cultivars of watermelon are highly diverse in fruit size (5–85 lb or 2.3–38.3 kg), shape (round, oval, oblong-cylindrical), rind color (very light to very dark green and often striped or mottled), flesh color (red, pink, orange, yellow, white), and seed size and color. The flesh contains 6–12% sugar, depending upon variety and condition of growth, with 8% sugar being acceptable on most markets, corresponding to juice refractometer readings of 9 or 10% total soluble solids. A few low-sugar cultivars called citrons are planted only for livestock. Such citron types grow wild in parts of Africa and are probably a significant factor in the survival of native animals in drought periods.

Seedless watermelons were introduced in 1948 in Japan. These are more difficult and costly to produce, yet they have become popular in some Asian countries. Seedless triploid watermelons are notably superior in flesh quality. Equally superior flesh quality occurs in some tetraploid varieties originated in the United States which are semiseedless and have extra firm flesh and hard rinds. Due to the extra stable flesh quality, these crops may be left on the vine for a single harvest.

The watermelon species is native to Africa, where it still occurs in the wild state in semiarid areas and is also widely cultivated by native farmers, both for human consumption and for livestock. Watermelons are grown in most tropical and subtropical and many temperate climates. They are grown locally in most of the United States, but in large volumes for interstate shipment primarily in the southern states and in California, Indiana, and Texas.

Food products. Watermelon juice of the sweet cultivars can be reduced to edible sugar and syrup. Watermelon rinds are either hard or soft, and this affects the handling performance of different cultivars. Varieties that best resist breakage in transit have firm interior flesh although the rind may be neither very thick nor very hard. Sweet pickles made from rinds have some commercial importance. Soft-rind varieties are best for that purpose.

Watermelon seeds are relished as food in some Near East countries and in China, and certain varieties are grown primarily for that purpose. There is a small industry in Iran where grilled (roasted) watermelon seeds are bagged and sold like popcorn. Salt-preserved watermelon seeds are often eaten in China.

Health value. Fresh watermelon flesh has a unique melting quality that has proved impossible to preserve in palatable form through any processing technique. The flesh consists of water (91%), fiber, and sugar, and little else of obvious nutritional value, but may have some as yet unproved health-promoting value. It is said to have diuretic properties, and both frozen concentrate and canned juice have been available for the treatment of nephritis. The seeds also are said to contain substances effective in the control of hypertension.

C. F. Andrus

Diseases. Over 35 diseases of watermelon have been described, but only 10 are of economic importance.

Soil-borne diseases. Fusarium wilt, caused by the fungus *Fusarium oxysporum* f. sp. *niveum*, occurs wherever watermelon is grown. The symptoms are wilting of the foliage and a brown discoloration of the vascular tissue in the taproot. *Fusarium* spp. and other fungi also cause damping-off, a disease of seedling watermelons resulting in their collapse and death. Root-knot nematodes (*Meloidogyne* spp.) are endoparasitic wormlike organisms that cause gall-like swelling on the roots.

Foliar diseases. Gummy stem blight and anthracnose occur on leaves, stems, and fruits. The fungus causing gummy stem blight, *Mycosphaerella citrullina*, produces large leaf lesions and is common on the stem and crown, while anthracnose, caused by the fungus *Glomerella cingulata* var. *orbiculare*, is particularly severe on the fruit (see **illus.**). Three fungal diseases restricted to the leaves are downy

Anthracnose lesions on a watermelon fruit.

mildew, caused by *Pseudoperonospora cubensis*, Cercospora leaf spot, caused by *Cercospora citrullina*, and Alternaria leaf spot, caused by *Alternaria cucumerina*. All three pathogens produce brown leaf spots with either large yellow borders (downy mildew), white lesion centers (Cercospora), or concentric rings of alternating light and dark brown tissues (Alternaria).

Watermelon mosaic is an aphid-transmitted disease caused by two viruses. Watermelon mosaic virus (WMV) 1 infects only cucurbits, while WMV 2 also infects legumes. Both produce mottling, distortion, and mosaic patterns on leaves and fruits.

Control. To control fusarium wilt and root-knot nematodes, a long-term rotation of 4–5 years between crops is recommended. A seed treatment with fungicidal chemicals is recommended to control damping-off. Most foliar diseases are controlled with fungicidal sprays that are applied to the leaves. Resistant cultivars are available for the control of fusarium wilt and anthracnose. There are no known economic controls for watermelon mosaic viruses. *See* PLANT PATHOLOGY.

N. C. Schenck

Bibliography. W. C. Adlerz, Spring aphid flights and incidence of watermelon mosaic 1 and 2 in Florida, *Phytopathology*, 64:350–353, 1974; D. L. Hopkins, Fungicidal control of downy mildew and gummy stem blight of watermelon, *Proc. Flor. State Hort. Soc.*, 85:108–110, 1972; O. C. Maloy, *Plant Disease Control: Principles and Practice*, 1993; G. K. Parris, *Diseases of Watermelon*, Flor. Agr. Exp. Sta. Bull. 491, 1952; N. C. Schenck, Watermelon disease incidence in central Florida, *Plant Dis. Rep.*, 44:556–558, 1960.

Waterpower

Power developed from movement of masses of water. Such movement is of two kinds: (1) the falling of streams through the force of gravity, and (2) the rising and falling of tides through lunar (and solar) gravitation.

While that part of solar energy expended to lift water vapor against Earth gravity is a minute fraction of the total, the absolute amount of energy that is theoretically recoverable from resulting streams is an enormous but unknown quantity. Of this, but a tiny portion is actually suitable for harnessing.

Water resources. The contribution of waterpower installations to the nation's electric power supply at the beginning of World War II was about 30%. While the output from hydroelectric plants has grown, their contribution to the total electric power has dropped to about 13%, because steam-electric plants have grown at a much more rapid rate.

As of March 1980, the Energy Information Agency of the Department of Energy estimated that by the end of 1993 the total developed capacity of waterpower installations could be 99,346 MW. However, that figure does not include a number of potential sites that could at some future time be considered. In theory, the Energy Information Agency estimates an eventual maximum potential of 187,000 MW, most of the undeveloped sites being in the Pacific Northwest and Alaska. Only a fraction of that will be developed for a variety of reasons. The most attractive sites have already been utilized. Hydro plants, with their initial high cost and generally long distances from major load centers, must compete with the large, efficient, fuel-fired stations, and the burgeoning, economical, large nuclear plants. Large dam sites usually must be justified not alone on the value of the power developed, but also on the benefits from flood control, irrigation, and recreation. Problems of migrating fish, conservation, and preservation of esthetic values are also factors. On the other hand, waterpower developments add greatly to power-system flexibility in meeting peak and emergency loads. Modern excavation and tunneling techniques are lowering construction costs. The economies of lowhead sites are improved by the new, efficient, axial-flow turbines of the tubular type. *See* ENERGY SOURCES.

Silting. The capacity of hydro plants cannot be counted on for perpetuity because of gradual filling of reservoirs with sediment. This effect is serious for irrigation, flood control, and navigation. Even when a lake behind a power dam becomes filled completely with silt, electric power can be generated on the run-of-the-river flow, although output would vary with stream flow.

The rate of silting varies widely with drainage basins. Because the Columbia River carries comparatively little silt, the reservoirs at Grand Coulee and Bonneville dams should have lives of many hundreds of years. The Colorado River, on the other hand, is muddy. In the first 13.7 years after Hoover Dam went into operation in 1935, 1424×10^6 acre-feet (175,600 hectare-meters) of silt was dumped into Lake Mead. That is equivalent to a layer 1 ft deep over 2225 mi^2 (or 1 m deep over 1756 km^2). This inflow of silt has been diminished about 22% by the construction of other dams upstream, for example, the Glen Canyon Dam. It is expected that Lake Mead will have a useful life of more than 500 years.

Pumped storage. In pumped-storage hydroelectric systems, water is pumped from a stream or lake to a reservoir at a higher elevation. Pumping up to a storage reservoir is most commonly done by reversing the hydraulic turbine and generator. The generator becomes a motor driving the turbine as a pump. Power is drawn from the power system at night or on weekends when demand is low. It is not practical to shut down large, high-temperature steam stations or nuclear units for a few hours at night or even over a weekend. Because they must run anyway, the cost of pumping power is low, whereas the power generated from pumped storage at peak periods is valuable. Also, the pumped-storage system provides a means of supplying power quickly in an emergency situation, for example, during the failure of a large

steam or nuclear unit. A pumped-storage system can be changed over from pumping to generation in 2 to 5 min. *See* PUMPED STORAGE.

Tidal power. A portion of the kinetic energy of the rotation of the Earth appears as ocean tides. The mean tide of all the oceans has been calculated as 2 ft (0.6 m), and the mean power as 5.4×10^{10} hp (40 TW) or, on a yearly basis, the equivalent of 3.6×10^3 kWh (4×10^{19} joules). Unfortunately, only a minute amount of this is likely to be harnessed for use. For tidal sites to be of sufficient engineering interest, the fall would have to be at least 15 ft (4.5 m). There are few such falls, and some of these are in remote areas. The only tide-power sites that have received serious attention are on the Severn River in England, the Rance River and Mont St. Michel in northern France, the San José and Deseado rivers of Argentina, the Petitocodiac and Memram-cook estuaries in the Bay of Fundy, Canada, the Passamaquoddy River where Maine joins New Brunswick, Canada, and the Cambridge Gulf of Western Australia.

The Passamaquoddy site, which has a potential of 1800 MW (peak), is the only important tidal-power prospect in the United States. However, engineers do not consider its electrical output to be economically competitive with power that is produced by other means.

A second major handicap to tidal power is that, with a simple, single-basin installation, power is available only when there is a several-feet difference between levels in the sea and the basin. Thus, firm power is not available. Also, periods of generation occur in consonance with the tide—not necessarily when power is needed.

The only major tidal power plant in operation is the one near the mouth of the Rance River in Normandy, France. This plant operates on 40-ft (12−m) tides. It began operation in 1967. It consists of twenty-four 10-MW bulb-type turbine-generator units of novel design. The system embodies a reservoir into which sea water is pumped during off-peak hours. Turbines are then run as pumps, power being drawn from the French electrical grid. The plant produces 5×10^8 kWh (1.8×10^{15} J) annually, including a significant amount of firm power.

Tidal power is an appealing and dramatic technique, and some other large plants may be constructed. However, the total contribution of the tides to the world's energy supply will be miniscule. *See* HYDROLOGY; TIDAL POWER. Charles A. Scarlott

Application. The basic relation for power P in kilowatt from a hydrosite is $P = QH/11.1$, where Q is water flow in ft^3/s, and H is head in feet. Actual power will be less as occasioned by inefficiencies such as (1) hydraulic losses in conduit and turbines; (2) mechanical losses in bearings; and (3) electrical losses in generators, station use, and transmission. Overall efficiency is always high, usually in excess of 80% to the station bus bars.

Choice of site. The competitive position of a hydro project must be judged by the cost and reliability of the output at the point of use or market. In most hydro developments, the bulk of the investment is in structures for the collection, control, regulation, and disposal of the water. Electrical transmission frequently adds a substantial financial burden because of remoteness of the hydrosite from the market. The incremental cost for waterwheels, generators, switches, yard, transformers, and water conduit is often a smaller fraction of the total investment than is the cost for the basic structures, real estate, and transmission facilities. Long life is characteristic of hydroelectric installations, and the annual carrying charges of 6–12% on the investment are a minimum for the power field. Operating and maintenance costs are lower than for other types of generating stations.

The fundamental elements of potential power, as given in the equation above, are runoff Q and head H. Despite the apparent basic simplicities of the relation, the technical and economic development of a hydrosite is a complex problem. No two sites are alike, so that the opportunity for standardization of structures and equipment is nearly nonexistent. The head would appear to be a simple surveying problem based largely on topography. However, geologic conditions, as revealed by core drillings, can eliminate an otherwise economically desirable site. Runoff is complicated, especially when records of flow are inadequate. Hydrology is basic to an understanding of water flow and its variations. Runoff must be related to precipitation and to the disposal of precipitation. It is vitally influenced by climatic conditions, seasonal changes, temperature and humidity of the atmosphere, meteorological phenomena, character of the watershed, infiltration, seepage, evaporation, percolation, and transpiration. Hydrographic data are essential in order to show the variations of runoff over a period of many years. Reservoirs, by providing storage, reduce the extremes of flow variation, which are often as high as 100 to 1 or occasionally 1000 to 1.

Economic factors. The economic factors affecting the capacity to be installed, which must be evaluated on any project, include load requirements, runoff, head, development cost, operating cost, value of output, alternative methods of generation, flood control, navigation, rights of other industries on the stream (such as fishing and lumbering), and national defense. Some of these factors are components of multipurpose developments with their attendant problems in the proper allocation of costs to the several purposes. The prevalence of government construction, ownership, and operation, with its subsidized financial formulas which are so different from those for investor-owned projects, further complicates economic evaluation. Many people and groups are parties of interest in the harnessing of hydrosites, and stringent government regulations prevail, including those of the U.S. Corps of Engineers, the Federal Power Commission, the Bureau of Reclamation, the Geological Survey, and the Securities and Exchange Commission.

Capacity. Prime capacity is that which is continuously available. Firm capacity is much larger and is dependent upon interconnection with other power plants and the extent to which load curves permit variable-capacity operation. The incremental cost for additional turbine-generator capacity is small, so that many alternatives for economic development of a site must be considered. The alternatives include a wide variety of base load, peak load, run-of-river, and pumped-storage plants. All are concerned with fitting installed capacity, runoff, and storage to the load curve of the power system and to give minimum cost over the life of the installation. In this evaluation it is essential clearly to distinguish capacity (kW) from energy (kWh) as they are not interchangeable. In any practical evaluation of water power in this electrical era, it should be recognized that the most favorable economics will be found with an interconnected electric system where the different methods of generating power are complementary as well as competitive.

As noted above, there is an increasing tendency in many areas to allocate hydro capacity to peaking service and to foster pumped-water storage for the same objective. Pumped storage, to be practical, requires the use of two reservoirs for the storage of water—one reservoir at considerably higher elevation, say, 500 to 1000 ft (150 to 300 m). A reversible pump-turbine operates alternatively (1) to raise water from the lower to the upper reservoir during off-peak periods, and (2) to generate power during peak-load periods by letting the water flow in the opposite direction through the turbine. Proximity of favorable sites on an interconnected electrical transmission system reduces the investment burden. Under such circumstances the return of 2 kWh on-peak for 3 kWh pumping off-peak has proven to be an attractive method of economically utilizing interconnected fossil-fuel, nuclear-fuel, and hydro power plants. *See* ELECTRIC POWER GENERATION; HYDRAULIC TURBINE; POWER PLANT; NUCLEAR REACTOR.

Theodore Baumeister

Bibliography. American Society of Civil Engineers, *Civil Engineering Guidelines for Planning & Designing Hydroelectric Developments*, vol. 5: *Pumped Storage and Tidal Power*, 1989; Annual statistical report for 1980, *Elec. World*, 193(6):49–80, March 15, 1980; E. A. Avallone and T. Baumeister III (eds.), *Marks' Standard Handbook for Mechanical Engineers*, 10th ed., 1996; B. W. Clowes (ed.), *Waterpower Eighty-seven*, 3 vols., 1988; A. J. Eberhardt (ed.), *Waterpower, 1989*, 3 vols., 1989; Edison Electric Institute, *Statistical Year Book of the Electric Utility Industry*, annually; D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 1999; M. G. Jog, *Hydroelectric and Pumped Storage*, 1989; D. Marier and L. Stoikaen (eds.), *Alternative Sources of Energy-Hydro*, No. 76, 1985; L. Moniton, M. LeNir, and J. Roux, *Micro Hydro-Electric Power Stations*, 1985; J. Raabe, *Hydro Power*, 1984; M. J. Roluti (ed.), *Waterpower, Nineteen-Eighty-Five*, 3 vols., 1986; *Water for Energy: Proceedings of the 3d International Symposium on Wave, Tidal, OTEC, & Small Scale Hydro Energy*, 1986.

Waterspout

An intense columnar vortex (not necessarily containing a funnel-shaped cloud) of small horizontal extent, over water. Typical visible vortex diameters are of the order of 33 ft (10 m), but a few large waterspouts may exceed 330 ft (100 m) across. In the case of Florida waterspouts, only rarely does the visible funnel extend from parent cloudbase to sea surface. Like the tornado, most of the visible funnel is condensate. Therefore, the extension of the funnel cloud downward depends upon the distribution of ambient water vapor, ambient temperature, and pressure drop due to the vortex circulation strength.

These vortices are most frequently observed during the warm season in the oceanic tropics and subtropics. In particular, the probable world maximum frequency for waterspout activity is over the Florida Keys. Another very prolific waterspout region is the entire southeast Florida Coast, although Tampa Bay has had the greatest number of damaging waterspouts. During the 1969 Lower Florida Keys Waterspout Project, nearly 500 separate waterspout events were documented for the period May to September.

All waterspouts undergo a regular life cycle composed of five discrete but overlapping stages.

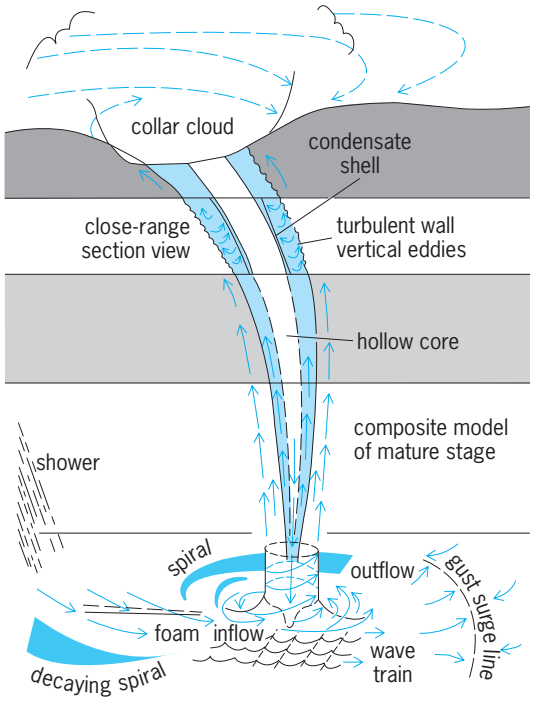


Fig. 1. Composite schematic model of a mature waterspout. For scaling reference, the maximum funnel diameters in this stage, just below the collar cloud, range from 10 to 460 ft (3 to 140 m).

Fig. 2. Decaying waterspout (in life-cycle stage 5) with the visible funnel retracting from the sea surface upward to parent cloudbase.

(1) The dark-spot stage is characterized by a prominent light-colored disk on the sea surface, surrounded by a dark patch diffuse on its outer edges—the dark spot may or may not have a small funnel above it initially, but signifies a complete vortex column extending from cloud-base to sea surface. (2) The spiral-pattern stage, the primary growth phase of the waterspout, is characterized by development of alternating dark- and light-colored bands spiraling around the dark spot on the sea surface. (3) The spray ring (incipient spray vortex) stage is characterized by a concentrated spray ring around the dark spot, with a lengthening funnel cloud above. (4) The mature waterspout stage (**Fig. 1**) is characterized by a spray vortex of maximum intensity and organization, the gradual weakening of the spiral pattern, and maximum funnel cloud length and diameter. (5) The decay stage occurs when the waterspout dissipates (often abruptly) as it is intercepted by the cool downdrafts from a nearby rain shower (**Fig. 2**).

Not every waterspout observed from its inception evolved through all other stages; however, the dark-spot stage and the mature waterspout stage have the greatest range of duration and often make up the bulk of a waterspout's lifetime (8–15 min average). In general, the larger-diameter waterspouts tend to be longer-lived and more intense; indeed, for south Florida waterspouts, about 10% or more each year equal or exceed tornadic size and intensity (maximum windspeeds just outside the calm "eye" of the spray vortex above 165 ft/s or 50 m/s). On the other hand, only about 1 out of 10 waterspouts in this region rotates anticyclonically (clockwise). Most waterspouts are spawned by cloudlines, which initially have average cloudtops of only 12,000–15,000 ft (3500–4500 m) in the Florida Keys. Waterspouts and tornadoes are qualitatively similar, differing only in certain quantitative aspects: tornadoes are usually more intense, move faster, and have longer lifetimes—especially maxi-tornadoes. Tornadoes are associated with intense, baroclinic (frontal), synoptic-scale disturbances with attendant strong vertical wind shear, while waterspouts are associated with weak, quasibarotropic disturbances (weak thermal gradients) and consequent weak vertical wind shear. *See* TORNADO; WIND.

Joseph H. Golden

Bibliography. J. H. Golden, An assessment of waterspout frequencies along the U.S. East and Gulf Coasts, *J. Appl. Meteorol.*, 16:231–236, 1977; J. H. Golden, The life-cycle of the Florida Keys waterspout, I, *J. Appl. Meteorol.*, 13:676–692, 1974; J. H. Golden, The lower Florida Keys waterspout project, May–September 1969, *Bull. Amer. Meteorol. Soc.*, 51:235–236, 1970; J. H. Golden, Scale interaction implications for the waterspout life cycle, II, *J. Appl. Meteorol.*, 13:676–692, 1974; J. H. Golden, Some statistical aspects of waterspout formation, *Weatherwise*, 26:108–117, 1973; J. H. Golden and D. Purcell, Life cycle of the Union City, Oklahoma, tornado and comparison with waterspouts, *Mon. Weath. Rev.*, 106(1):3–11, 1978.

Watt balance

An electromechanical apparatus for establishing the watt as an SI electrical unit. Prior to January 1, 1990, there were voltage units in use in the United States, France, the United Kingdom, Russia, and so on which differed from each other (and, with hindsight, from the SI volt) by up to 9 parts per million (ppm). These units were based on results from various current balances. Then results obtained from a different kind of apparatus at the National Physical Laboratory in the United Kingdom and the National Institute for Science and Technology in the United States, which were based on the simple principle described below, enabled the electrical units to be put on a sound SI basis. The accuracy in deriving the SI unit of voltage in this way was considered to be better than 0.2 ppm. Since that date these apparatuses have continued to be refined, and others are being developed in metrological laboratories, with the more ambitious objective of defining the kilogram in terms of fundamental physical constants instead of having to rely on a carefully preserved artifact, the cylinder of platinum-iridium alloy kept at the International Bureau of Weights and Measures. To do this, the accuracy achieved by the apparatus will have to be demonstrated to be of the order of 0.01 ppm. *See* CURRENT BALANCE; ELECTRICAL UNITS AND STANDARDS.

Working principle. A conductor whose total length is l is wound into a horizontal circular coil and placed in a horizontal magnetic flux B which is everywhere perpendicular to the coil (see **illustration**). If a current I flows in the conductor, the consequent vertical force can be opposed by a mass M which is subject to the earth's gravitational acceleration, g. That is, Eq. (1) is satisfied.

$$Mg = BIl \tag{1}$$

See MAGNETISM.

If, in a separate measurement, the coil is moved vertically with velocity u, a voltage V will appear

across it and will be given by Eq. (2).

$$V = Blu \tag{2}$$

See ELECTROMAGNETIC INDUCTION; FARADAY'S LAW OF INDUCTION.

Combining Eqs. (1) and (2) eliminates the product Bl, which is too difficult to measure with the required accuracy, and gives Eq. (3). This equation

$$VI = Mgu \tag{3}$$

relates the product VI to Mgu watts as measured directly in terms of the SI units of mass, length, and time.

The ambition of the present effort is to modify the SI by defining voltage and current units in terms of the Josephson effect constant, $2e/h$, and the quantum Hall effect constant, h/e^2. The effect of this would be to redefine the mass unit, the kilogram, in terms of a defined value of Planck's constant, h, the second, and the velocity of light, which defines the meter. *See* HALL EFFECT; JOSEPHSON EFFECT.

Advantages of this approach. Equation (3) can be derived in a much more general manner to bring out two points vital to attaining the required accuracy:

1. Equation (3) contains no terms relating to actual expenditure of energy. That is, the relation is concerned with virtual work only, and no real work such as friction in the bearings of any balance used to judge equality of the electromagnetic and gravitational forces, or joule heating of the coil by the current, is relevant to its exactness.

2. The geomertical shape of the coil and the spatial distribution of the magnetic flux are also irrelevant. Circular and rectangular coils and linear and radial fluxes from permanent or superconducting magnets have been employed. All that matters is that coil and flux should combine to give a vertical force, a condition which can be tested in various experimental ways.

A further advantage of this approach to linking the electrical and base SI units is that the magnitude of the various quantities can easily be designed to be of a size suitable for measurement at the highest level of accuracy possible. To illustrate this, a kilogram weight can be opposed by a few milliamperes of current which can be measured in terms of the potential difference it produces across a 100-ohm resistor. A velocity of a few millimeters a second, which is measurable as a few thousand optical fringes a second by a laser interferometer, can produce a volt of emf which can be compared directly with the output of an array of Josephson junctions. Present-day technology can measure g to 0.001 ppm accuracy by interferometric distance measurement and timing of a freely falling object. Nevertheless, attaining an overall accuracy of the order of 0.01 ppm from watt balance measurements demands unprecedented skills and meticulous attention to detail in mechanical, optical, and electrical metrology. As well as creating the satisfaction of progressing towards replacing the last SI base unit based on an artifact, undertaking watt balance measurements advances capabilities in these areas of metrology. *See* PHYSICAL MEASUREMENT.

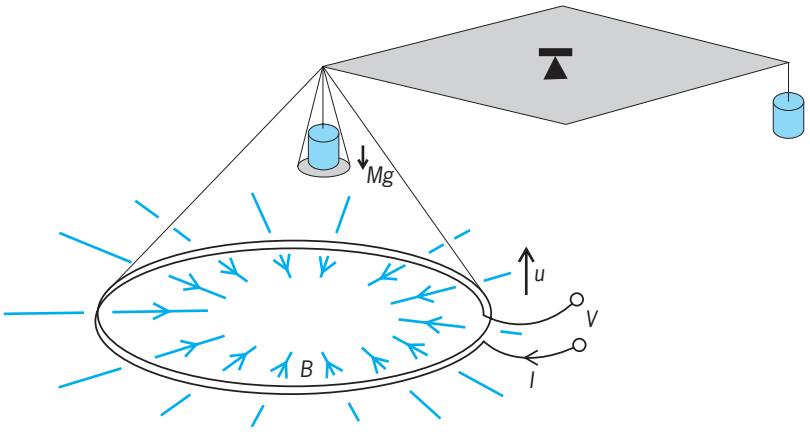


Diagram of the Watt balance, illustrating the working principle of the apparatus. Symbols are defined in text.

Bryan P. Kibble

Bibliography. A. Eichenberger et al., Tracing Planck's constant to the kilogram by electromechanical methods, *Metrologia*, 40:356–365, 2003; I. Robinson, Redefining the kilogram, *Phys. World*, 17(5):31–35, May 2004.

Watt-hour meter

An electrical energy meter, that is, an electricity meter that measures and registers the integral, with respect to time, of the power in the circuit in which it is connected. This instrument can be considered as having two parts: a transducer, which converts the power into a mechanical or electrical signal, and a counter, which integrates and displays the value of the total energy that has passed through the meter. Either or both of these parts can be based on mechanical or electronic principles.

In its wholly mechanical form the transducer is an electric motor designed so that its torque is proportional to the electric power in the circuit. The motor spindle carries a conducting disk that rotates between the poles of one or more strong permanent magnets. These provide a braking torque that is proportional to the disk rotational speed, so the motor runs at a rate that accurately represents the circuit power. The integrating register is simply connected to the motor through a gear train that gives the required movement of the dials in relation to the passage of electrical energy. *See* MOTOR.

Mechanical meters. These can measure either dc or ac energy.

DC energy measurement. Some form of commutator motor similar to those with a shunt field winding is commonly used for dc energy measurement. It is most convenient for the field to carry the circuit current, while the armature is fed with a signal from the circuit voltage. The motor is designed to have very

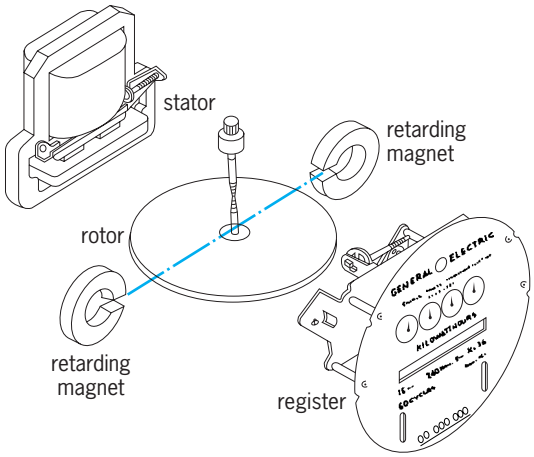


Fig. 1. Basic elements of an induction-type watt-hour meter. (*General Electric Co.*)

low friction and to start easily in order to maintain a linear torque-to-power relationship over as wide a range as possible. Some early designs were similar to the homopolar motor, in which the armature is a copper disk with the current entering through a pool of mercury. Owing to the wish to reduce mercury in the environment, because of its toxicity, these are no longer likely to be installed.

An entirely different approach to dc energy metering depends on the change in the rate of swing of a pendulum that carries a coil on its bob and swings over a second, fixed coil. The swinging coil carries a signal proportional to the voltage, while the fixed coil carries a signal depending on the load current. A duplicate system is arranged with one of the current directions reversed. The rate of one pendulum is increased by the power being measured, and the other decreased, as a result of the magnetic forces between the fixed and swinging coils. A pair of escapements is controlled by the pendulums, and the difference in their rates is obtained by using a mechanical differential. The rate of rotation of the differential output is proportional to the circuit power and is used to drive a mechanical register. The obvious error arising from mismatch of the pendulums is eliminated by regularly reversing the directions of the currents in the swinging coils and the direction of mechanical drive. These instruments are capable of 0.1% accuracy. *See* DIFFERENTIAL; PENDULUM.

AC energy measurement. The Ferraris, or induction-type, meter is used for ac energy measurement (**Fig. 1**). The stator (**Fig. 2**) carries two windings. The potential winding is fitted onto the single limb of the laminated iron structure. It is arranged to have a highly inductive impedance in order that the current flowing in it is in quadrature with the applied voltage. The current coils are wound on a pair of limbs situated on the other side of the structure. The rotor disk is mounted on a vertical axle with low-friction bearings and is arranged so that it passes through gaps in the stator and one or more permanent magnets. The two fluxes from the current and potential coils induce currents in the disk that interact with them to produce a torque proportional to the circuit power. The magnets produce a retarding torque that is proportional to the rotor speed, and is therefore proportional to the power. The disk drives a mechanical register through a worm drive on its axle. A typical rate of disk rotation is 200 revolutions per kilowatt-hour. An ordinary energy meter will easily achieve an accuracy of 2% over a wide range of loads; precision models may reach 0.1%.

Polyphase energy meters may be constructed with two or three stators driving the same disk, or individual disks mounted on the same axle. Care is required in these designs if interaction between the elements is to be avoided. Additional indicators can be fitted to indicate the maximum demand that has been made during, for example, any half-hour period since the instrument was reset. Other modifications enable the meter to record reactive energy, in kilovolt-ampere-reactive-hours, instead of real energy, in kilowatt-hours. Remote metering can be achieved by arranging for the rotation of the disk to provide electrical impulses through contact closures or optical rotation detectors. *See* VOLT-AMPERE.

Ferraris meters are in very wide use and measure the consumption of the vast majority of domestic and industrial users of electric power throughout the world. The design has been refined over many years, and its simplicity, combined with mass-production methods, has held manufacturing costs to a level at which it was hard for any other approach to compete. However, the demand on the one hand for much greater precision and on the other hand for greater complexity of measurement has led to the successful development of a variety of electronic watt-hour meters.

Electronic meters. The electronic watt transducer is a solid-state circuit that performs the multiplication of current and voltage signals, and delivers an output in the form of a pulse train at a rate proportional to power. The most widely used principle is the time-division multiplier, in which pulses are modulated in duration and amplitude by the voltage and current

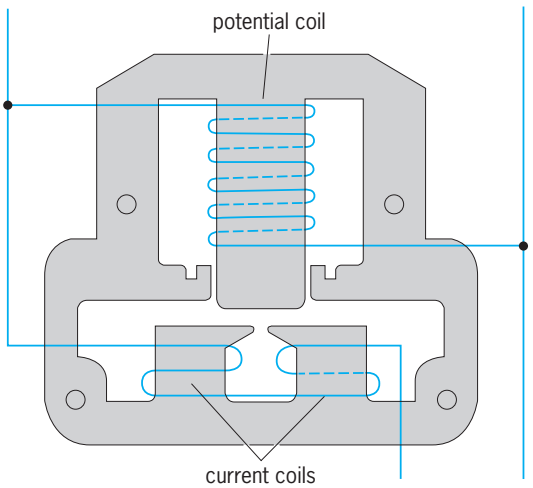


Fig. 2. Stator of induction-type watt-hour meter, showing the windings. (*General Electric Co.*)

signals, respectively. A suitably scaled version of the voltage signal is compared in a comparator with the output of a triangle or sawtooth waveform. The comparator output is a train of pulses whose durations are proportional to the quasi-instantaneous values of the voltage. In effect, this is a sampling system, and the pulse rate needs to be high compared with the signal frequency. Rates of 20–100 kHz are typical. The comparator output is then modulated to produce pulses having amplitudes proportional to the current signal. The resulting pulse train is then smoothed, giving a dc voltage signal that represents the power level. *See* COMPARATOR; PULSE MODULATION.

A further section converts this power signal into another train of pulses, but the pulse rate is now arranged to be proportional to the power. One method of achieving this is to apply the dc signal to an integrator. Each time the integrator output exceeds a threshold level, a single pulse is generated. This has its amplitude and duration very accurately defined and is applied to the integrator input with the polarity required to return the integrator toward zero. In the long term, pulses will be generated at the correct rate to be proportional to the input current regardless of the actual threshold level or integrator components.

The simplest solid-state watt-hour meter is completed by adding an electronic register to record the energy consumed. Precision electronic energy meters can give errors less than 0.005%. Electronic energy meters are not directly competitive with mechanical versions where a simple meter at the 2% level is required. However, for any application where more elaborate features are needed, the electronic model is superior and the absence of any visible moving parts reduces the opportunity for fraud. Electronic instruments are available in which six registers are provided, to record consumption at four different times of day and two levels of maximum demand. Automatic switching of heating loads can be built in, controlled by internal clocks, signals carried on the power lines, or radio signals. Prepayment and recovery of debts can be arranged through special meters that are used in conjunction with electronic keys, magnetic cards, or smart cards. Advanced designs are based on the independent measurement of voltage, current, and phase, with a microprocessor carrying out the calculation of any required parameter, including kilowatt-hours, kilovolt-ampere-reactive hours, and power factor. *See* MICROPROCESSOR.

Intelligent, or smart, meters can provide a wide variety of load and tariff-control functions, as well as remote reading of energy consumption. Different approaches are used around the world in order to meet local requirements and use available facilities. One-way or two-way communication may be operated, using radio, telephone, or the power line itself. In the United States, 900-MHz radio links have been used to enable a moving van to interrogate all the meters in a locality. In the United Kingdom, where this frequency band is used for other purposes, over 1.5 million meters are in use which select multiple tariffs and switch heating loads under the control of a radio teleswitch, operated by phase modulation of a broadcast transmission at 180 kHz. A different system uses dedicated transmissions in the 183.5–184.5-MHz band. In continental Europe, where many low-frequency systems have been in use for a long period, signals are superimposed on the power network over a low-frequency band at 9–95 kHz. Although the use of the network is very attractive, being under the control of the corporation directly concerned and saving payments to third parties, there are serious difficulties in its use owing to the way in which its transmission characteristics change with load. Broad-spectrum transmissions are employed, and reliable communication at 200 baud is achieved.

Hybrid and special meters. Electronic and mechanical techniques can be combined in a variety of ways. Signals from a mechanical transducer may be used to operate electronic registers in order to obtain the advantages of the facilities that they can provide. A mechanical impulse register may be used in conjunction with an electronic transducer, where it is considered important to maintain a record without the need for batteries or nonvolatile memory elements.

Watt-hour meters that operate at potentials of 100–250 V and currents up to 100 A are widely manufactured. At higher levels, voltage or current transformers are used to reduce the signals handled by the meter to more convenient values, frequently 110 V and 5 A. By this means it is possible to carry out energy metering at any level required, including hundreds of kilovolts and tens of kiloamperes. *See* ELECTRICAL ENERGY MEASUREMENT; ELECTRIC POWER MEASUREMENT; INSTRUMENT TRANSFORMER; TRANSFORMER. R. B. D. Knight

Bibliography. T. Brewer, *An Introduction to Electrical Measurements*, 2d ed., 2001; *Electric Meters Code for Electricity Metering*, Amer. Nat. Stand. C-12.1, 2001; R. Dettmer, Meter power, *IEE Rev.*,38(6):237–240, 1992; Edison Electric Institute, *Handbook for Electricity Metering*, 9th ed., 1999; P. S. Filipski, A TDM wattmeter with 0.5-MHz carrier frequency, *IEEE Trans. Instrum. Meas.*, IM-39:15–18, 1990; F. K. Harris, *Electrical Measurements*, 1952, reprint 1975; M. V. Reissland, *Electrical Measurements*, 1989; *Watthour Meters*, Amer. Nat. Stand. C-12.10, 1987.

Wattmeter

An instrument that measures electric power. For a complete discussion of power measurement in various types of electric circuits. *See* ELECTRIC POWER MEASUREMENT.

A variety of wattmeters are available to measure the power in ac circuits. They are generally classified by names descriptive of their operating principles. Determination of power in dc circuits is almost always done by separate measurements of voltage and current. However, some of the instruments described will also function in dc circuits, if desired.

Electrodynamic wattmeter. Probably the most useful instrument in the measurement of ac power at

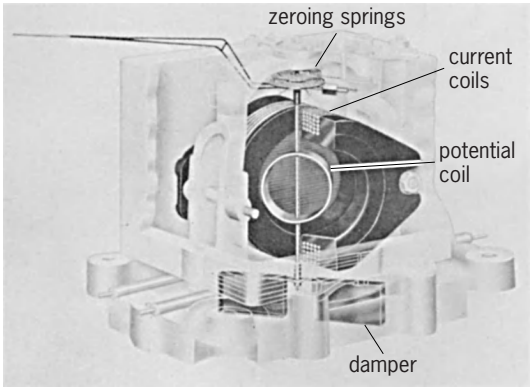


Fig. 1. Single-element electrodynamic wattmeter. (*Weston Instruments, Division of Daystrom, Inc.*)

commercial frequencies is the indicating (deflecting) electrodynamic wattmeter (**Fig. 1**). It is similar in principle to the double-coil dc ammeter or voltmeter in that it depends on the interaction of the fields of two sets of coils, one fixed and the other movable. The moving coil is suspended, or pivoted, so that it is free to rotate through a limited angle about an axis perpendicular to that of the fixed coils. As a single-phase wattmeter, the moving (potential) coil, usually constructed of fine wire, carries a current proportional to the voltage applied to the measured circuit, and the fixed (current) coils carry the load current. This arrangement of coils is due to the practical necessity of designing current coils of relatively heavy conductors to carry large values of current. The potential coil can be lighter because the operating current is limited to low values. *See* AMMETER.

If there is no iron or disturbing magnetic field due to current in neighboring conductors, then the instantaneous torque is proportional to the product of the instantaneous current in the potential coil and the instantaneous current in the current coils. Since the instantaneous current in the potential coil is proportional to the instantaneous voltage across the circuit, it follows that the instantaneous torque is proportional to the product of the instantaneous voltage and the instantaneous current in the current coils.

The moving system, however, is designed with sufficient inertia that it is unable to follow the rapid alternations of the alterating current, but it will rotate, opposed by light zeroing springs, to a position corresponding to the average torque, which is proportional to the average power being supplied.

To avoid a loss of power in the instrument due to the current flowing in the potential coil, a fine-wire coil is wound with the current coils and connected in series with the potential coil. Its effect is to cancel the magnetic effect of the potential coil current from the field of the current coils. If this compensation is correct, with the load circuit open, the instrument will read zero.

The presence of inductance in the potential circuit would normally introduce phase displacement between voltage and current, which is theoretically inadmissible. Simple tuning with capacity would introduce frequency error, but capacity in series with the moving coil and shunted by noninductive resistance is found to reduce the net reactance to a tolerably low value. *See* ALTERNATING-CURRENT CIRCUIT THEORY; CAPACITANCE; INDUCTANCE; REACTANCE.

Other disturbing effects are those of ambient temperature, eddy currents in metal structures close to the moving coil, transformer effect due to mutual induction between current and potential coils, and skin effect in which a change in distribution of current in the current coil causes frequency error. In precision instruments these effects are taken care of as much as possible by proper choice of materials and by shielding. *See* EDDY CURRENT; SKIN EFFECT (ELECTRICITY); TRANSFORMER.

This type of wattmeter is capable of great accuracy within its range and is commonly used as a laboratory standard. It has the great advantage of being usable as a transfer instrument; that is, it can be calibrated with direct current and used for ac measurements. It is used chiefly at power frequencies. For reasons of manufacturing convenience rather than on technical grounds, commercially available instruments are limited to full scale ranges of 300 V and 10 A. External resistive multipliers may be used to extend voltage range. Shunting of current coils is not recommended. The preferred method of extending the range of both current and voltage is through current and potential transformers, by which the voltages and current in the load circuit are reduced to nominal values by definite transformation ratios without introducing appreciable phase errors, or, if such errors are introduced, their values are known and may be accounted for. *See* INSTRUMENT TRANSFORMER.

Quarter-squares multiplying wattmeter. The mathematical relationship of Eq. (1) can be used to obtain

$$4XY = (X + Y)^2 - (X - Y)^2 \qquad (1)$$

the effect of an analog multiplier. If signals representing the voltage (X) and current (Y) are combined to form new signals equal to the sum and difference, and these signals are then squared and the difference of the squares is evaluated, the resulting output is proportional to the required product. Equation (1) shows that a scale factor of 4 is introduced, which gives the method its name. Several different wattmeters are based on this principle.

Thermal wattmeter. A thermal converter consists of a resistive heater in close thermal contact with one or more thermocouples. When current flows through the heater, the temperature rises. Thermocouples give an output voltage proportional to the temperature difference between their junctions, in this case proportional to the square of the current, and so make suitable transducers for the construction of thermal wattmeters. *See* ELECTRICAL RESISTANCE.

In one possible arrangement (**Fig. 2**), four devices are used in a bridge configuration. The original voltage and current signals are first converted into currents within the working range of the thermal converters. The sum of these two currents flows through two of the devices, while the difference flows through the other two devices. The

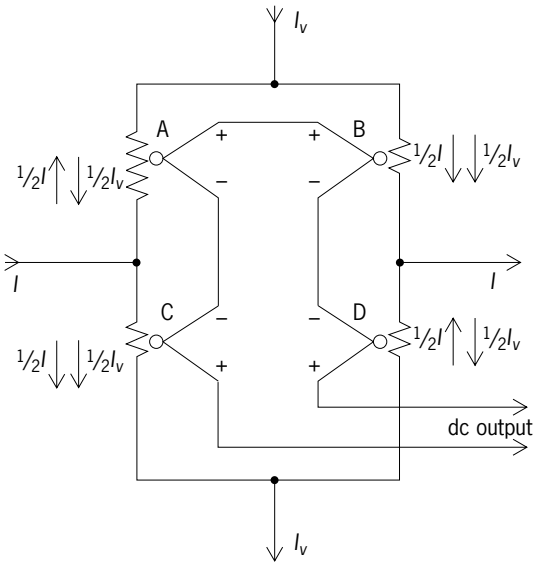


Fig. 2. Thermal wattmeter circuit. I_v is the voltage signal; I is the current signal; B and C are thermal converters with current $^1/_2(I + I_v)$; A and D are thermal converters with current $^1/_2(I - I_v)$.

required difference between the dc outputs is obtained by suitable connection of the four thermocouples. Close matching of the thermal converters is necessary in order to ensure that the subtraction accurately eliminates the signals corresponding to the squares (X^2 and Y^2) of the voltage and current inputs. Instruments of this type can have a wide bandwidth, owing to the simplicity of the thermal converter, but accuracy is limited to a few percent.

Considerable improvements in accuracy result from the use of a single multijunction thermal converter switched sequentially between the sum and a second signal obtained by combining the difference with a third, dc signal. This dc signal is derived by a feedback arrangement designed to cause the two inputs to become equal. Rearranging Eq. (1) shows that this dc signal gives the required power measurement. Wattmeters of this design can have errors as low as 10 parts per million at power frequencies. *See* THERMAL CONVERTERS; THERMOCOUPLE; THERMOELECTRICITY.

Electrostatic wattmeter. The electrostatic force between two conductors is proportional to the product of the square of the potential difference between them and the rate of change of capacitance with displacement. A differential electrostatic instrument may therefore be used to construct a quarter-squares wattmeter. In spite of the problems of matching the capacitance changes of the two elements and the small forces available, electrostatic wattmeters were used as standards for many years. Uncertainties of the order of 50 parts per million could be obtained with a useful frequency response up to 100 kHz, but the working levels of current and voltage were very limited.

Digital wattmeter. Digital wattmeters combine the advantages of electronic signal processing and a high-resolution, easily read display. Electrical readout of the measurement is also possible. A variety of electronic techniques for carrying out the necessary multiplication of the signals representing the current and voltage have been used. Usually the electronic multiplier is an analog system which gives as its output a voltage proportional to the power indication required. This voltage is then converted into digital form in one of the standard ways. Many of the multipliers were originally developed for use in analog computers. *See* ANALOG COMPUTER; ELECTRONIC DISPLAY.

Hall-effect multiplier. The Hall-effect multiplier makes use of the relationship between the transverse (Hall) voltage which appears across a suitable semiconductor device, the current flowing through it, and an applied magnetic field. The voltage is proportional to the product of the device current and the magnetic field. Multiplication can be performed in all four quadrants (that is, when the current and field are, respectively, both positive, positive and negative, negative and positive, or both negative). This system can give wide bandwidth, but accuracy is limited by the small output from the device and its poor drift characteristics. *See* HALL EFFECT.

Pulse height/width multiplier. The pulse height/width multiplier is in widespread use. In one version the width of pulses in a train having constant frequency is arranged to be controlled by one signal, generally the voltage, while their height is controlled by the other. The resulting signal is smoothed by a filter, which gives an output proportional to the required product. Some ingenuity is required to extend the method to work in four quadrants, and the bandwidth is limited by the need for the pulse frequency to be higher than the highest-frequency component of the power to be measured. At 60 Hz, errors as low as 10 parts per million can be realized. *See* ELECTRIC FILTER; PULSE MODULATION.

Electronic quarter-squares wattmeter. Electronic versions of the quarter-squares multiplier have been developed. The combination of the voltage and current signals may be carried out by using operational amplifiers. As described above, thermal instruments can achieve very high accuracy. Instruments have also been made which use electronic squaring circuits, similar to those developed for root-mean-square sensing. In such instruments, the conventional expression for power is replaced by the difference between the average values of the squares in Eq. (1) over long time periods, the evaluation of these two squares being carried out separately. The relative phase of the sum and difference signals is not significant, unlike the relative phase between the voltage and current signals, which is critical to the measurement. The technique is attractive for high-frequency applications, where minimizing phase errors is a major design problem. *See* OPERATIONAL AMPLIFIER.

A variant of the electronic quarter-squares multiplier (**Fig. 3**) makes use of the relationship in Eq. (2).

$$2XY = (X + Y)^2 - X^2 - Y^2 \qquad (2)$$

This simplifies the phase problems further, since only

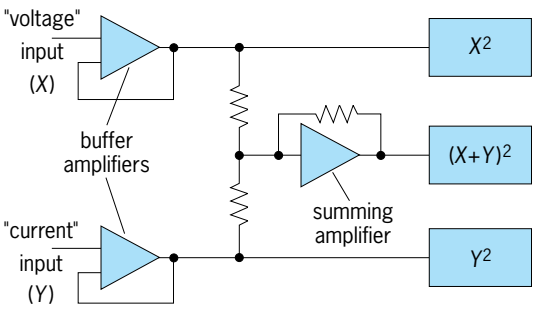


Fig. 3. Wideband electronic wattmeter.

the sum needs to be produced. Since the measurements of the three terms are independent, X^2,Y^2, and $(X + Y)^2$ can be evaluated by using conventional high-accuracy digital voltmeters, with the square being obtained arithmetically from the root-mean-square value indicated. This technique has been applied to an audio-frequency power standard which has errors of less than 50 parts per million at 3 kHz and 200 parts per million at 30 kHz.

Digital sampling wattmeter. A different approach uses sampling techniques (**Fig. 4**). Sets of coincident samples of the voltage and current signals are obtained, using sample-and-hold units. The resulting instantaneous values are converted into digital form by fast analog-to-digital converters. The pairs of measurements are multiplied by using a computer, which accumulates the products to provide a value for the power. A steady result can be obtained by accumulating samples over a time (and number of samples) which is long compared with the period of the waveform. Another strategy is to carefully control the timing of the samples so that N samples are taken in exactly M cycles, where N and M have no common factors. This has been shown to be a very powerful method, as the response of instruments using the method can extend to frequencies limited by the sample-and-hold units rather than the sampling rate. Other arrangements can include the use of random sample timing or interpolation between samples. The accuracy is limited by the performances of available sample-and-hold units and analog-to-digital converters, which are approaching the performances of thermal devices. *See* ANALOG-TO-DIGITAL CONVERTER.

Polyphase wattmeter. The instruments thus far considered are designed for single-phase power measurement. In polyphase circuits, the total power is the algebraic sum of the power in each phase. This summation is assisted by simple modifications of single-phase instruments. *See* ALTERNATING CURRENT.

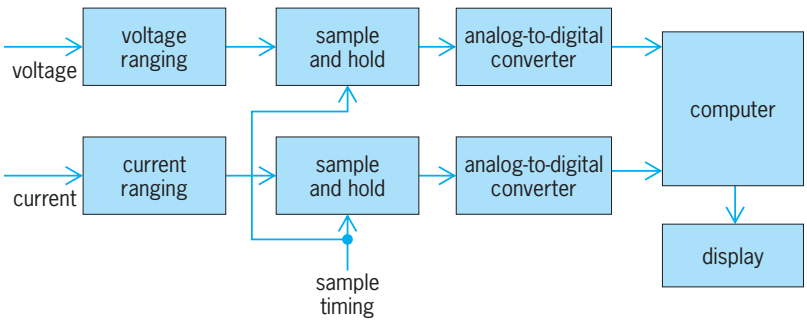


Fig. 4. Block diagram of digital sampling wattmeter.

For example, an electrodynamic wattmeter may contain a second coil system similar to the coils of a single-phase meter, with the second potential coil on the same shaft as the potential coil in the first system. The two systems are mounted in the same case and are designed to have matched characteristics, but care is taken that there is no magnetic interaction between them. The deflection is proportional to the sum of the torques of the two elements; thus, total power is read from the instrument scale. Electrical connections are the same as for two single-phase wattmeters. Electronic and other wattmeters in which the measurement of power appears as a voltage or digital signal may be extended to polyphase applications by suitably combining the outputs from a number of single-phase transducers.

Loading effects. When a wattmeter is used in real measurement situations, there is a problem in connecting the instrument without introducing a source of error, as the power consumed in either the voltage- or current-measuring circuit will be included in the measured power. If the voltage circuit is connected directly across the load, the current circuit will sense the current consumed by it. Conversely, if the voltage circuit is connected on the supply side of the current sensor, it will include the voltage drop in the current circuit in the measured voltage. As most wattmeters are used at a voltage close to a nominal value, the errors caused by losses in the voltage circuit are almost constant. It is therefore possible to design instruments which incorporate compensation for these losses, and the most commonly used connection applies the voltage sensing circuit directly across the load.

Range extension. The full-scale ranges of voltage and current of all types of instrument are extended by the use of electronic amplifiers, voltage attenuators, voltage transformers, or current transformers. Such scaling devices commonly bring the signals to the nominal level of 110 V, 5 A. In this way, power at extremely high voltages and currents can be measured, with all connections to the wattmeter being in light, modestly insulated cables. *See* AMPLIFIER.

Calibration. When wattmeters are being calibrated, it is not necessary to operate the test and standard instruments in a situation where real power exists. If the voltage circuits are connected in parallel and the current inputs are connected in series, both instruments are deflected by the same product of voltage and current, although only sufficient power is consumed as is required to operate the measurement circuits. This technique is known as phantom power calibration. When the wattmeter is operated in this way, loading effects do not appear. In order to ensure accurate calibration, instruments providing compensation of loading effect usually make

provision for the compensation to be switched off in these circumstances. R. B. D. Knight

Bibliography. F. K. Harris, *Electrical Measurements*, 1952, reprint 1975; A. D. Helfrick and W. D. Cooper, *Modern Electronic Instrumentation and Measurement Techniques*, 1990; M. V. Reissland, *Electrical Measurements: Fundamentals, Concepts, Applications*, 1989; M. G. Say (ed.), *Electrical Engineer's Reference Book*, 15th ed., 1993; L. Schnell (ed.), *Technology of Electrical Measurements*, 1993; L. M. Thompson, *Electrical Measurements and Calibration*, 2d ed., 1994.

Wave (physics)

The general term applied to the description of a disturbance which propagates from one point in a medium to other points without giving the medium as a whole any permanent displacement.

Waves are generally described in terms of their amplitude, and how the amplitude varies with both space and time. The actual description of the wave amplitude involves a solution of the wave equation and the particular boundary conditions for the case being studied. In the cases most often considered, the wave equation is simplified to a second-order, linear, partial differential equation. This equation for a one-dimensional space coordinate x is written as Eq. (1), where ϕ is the amplitude, c is the wave velocity, and t is the time coordinate. The generalized solutions of this equation are of the form given by Eq. (2). The first term indicates a wave traveling in the positive x direction at a velocity c, and the second term a wave traveling in the negative x direction. The functions F and G are determined by the particular properties of the boundary conditions of the problem. In the one-dimensional case these are usually sine or cosine waves. The velocity of propagation c is proportional to the square root of the ratio of the elastic to the inertial constants of the medium. *See* SINE WAVE; WAVE EQUATION; WAVE MOTION.

$$\frac{\partial^2\phi}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} \tag{1}$$

$$\phi = F(x - ct) + G(x + ct) \tag{2}$$

Acoustic waves, or sound waves, are a particular kind of the general class of elastic waves. Elastic waves are propagated in media having two properties, inertia and elasticity. Elasticity of the medium is required in order to provide a force which tends to restore a displaced particle of the medium to its original position. Inertia is required to enable the displaced particle to transfer momentum to an adjoining particle. A shear wave, or rotational wave, is a wave in an elastic medium which causes an element of the medium to change its shape without a change of volume.

Electromagnetic waves (for example, light waves and radio waves) are not elastic waves and therefore can travel through a vacuum. The velocity of the wave depends on the medium through which the wave travels, but in a vacuum it is a constant, c, approximately equal to 3×10^8 m/s (9.8×10^8 ft/s). *See* ELECTROMAGNETIC WAVE; SHOCK WAVE.

William J. Galloway

Wave equation

The name given to certain partial differential equations in classical and quantum physics which relate the spatial and time dependence of physical functions. In this article the classical and quantum wave equations are discussed separately, with the classical equations first for historical reasons.

Classical wave equation. In classical physics the name wave equation is given to the linear, homogeneous partial differential equations which have the form of Eq. (1). Here v is a parameter with the dimensions of velocity; $\mathbf{r}$ represents the space coordinates x, y, z; t is the time; and ∇^2 is Laplace's operator defined by Eq. (2). The function $f(\mathbf{r}, t)$ is a physical observable; that is, it can be measured and consequently must be a real function. Laplace's operator may also be written in polar coordinates $\mathbf{r}$, θ, and ϕ; if f depends only on the radial distance r, the wave equation has the simple form shown by Eq. (3).

$$\left[\nabla^2 - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right] f(\mathbf{r}, t) = 0 \tag{1}$$

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{2}$$

$$\left[\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} - \frac{1}{v^2}\frac{\partial^2}{\partial t^2}\right] f(r, t) = 0 \tag{3}$$

Motion of a string. The simplest example of a wave equation in classical physics is that governing the transverse motion of a string under tension and constrained to move in a plane. In this case the wave equation is given by Eq. (4), where $y(x, t)$ is the transverse displacement of the string from its equilibrium position at time t. If the string is thought of as being made up of many separate particles, x is the label for the particle. The velocity parameter is $v = \sqrt{T/\lambda}$, with λ the linear mass density and T the tension in the string, both uniform. Equation (4) is valid when the slope of the string is small, that is, $\partial y/\partial x \ll 1$. A solution to Eq. (4) is any function of the form $f(x - vt)$, where f is arbitrary except that it must be twice differentiable. Since at $t = 0$ one would have $y(x, 0) = f(x)$, $f(x)$ specifies the initial shape of the string. Since $f(x - vt)$ differs from $f(x)$ only by a translation along the x axis of magnitude vt, it is apparent that the shape of the string moves along the x axis as a wave with shape determined by $f(x)$ and with velocity v, which is called the wave propagation velocity. A point to note is that v is not the transverse velocity $\partial y/\partial t$ of the particles in the string. In fact,

$$\frac{\partial^2 y(x, t)}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 y(x, t)}{\partial t^2} \tag{4}$$

since Eq. (5) holds, the transverse velocity for the

$$\frac{\partial y}{\partial t} = -v\frac{\partial y}{\partial x} \tag{5}$$

solution $f(x - vt)$ is always small and opposite in sign to the slope. Another solution to Eq. (4) could be $g(x + ct)$. which corresponds to the shape of the string moving in the negative x direction, that is, with velocity $-v$. The most general solution to Eq. (4) is the linear combination given by Eq. (6).

$$y(x, t) = a^{+}f(x - vt) + a^{-}g(x + vt) \tag{6}$$

The particular choice of constants $a\pm$ is determined by the shape of the string at $t = 0$ and by the transverse velocity at $t = 0$, that is, by $y(x, 0)$ and $\partial y(x, t)/\partial t|_{t=0}$.

Acoustical waves. A second type of classical physical situation in which the wave equation, Eq. (1), supplies a mathematical description of the physical reality is the propagation of pressure waves in a fluid medium. Such waves are called acoustical waves, the propagation of sound being an example. Suppose that x is the undisturbed position of a particle in the fluid and let the displacement at time t be $z(x, t)$ so that those particles that have position x when undisturbed have position $x + z(x, t)$ at time t after being disturbed. If $\partial z/\partial x \ll 1$, the acoustic approximation, then the displacement $z(x, t)$ satisfies Eq. (7), where

$$\frac{\partial^2 z(x, t)}{\partial x^2} = \frac{1}{\overline{v}^2}\frac{\partial^2 z(x, t)}{\partial t^2} \tag{7}$$

$\overline{v}^2 = dP(\rho)/d\rho$ at $\rho = \rho_0$, with ρ the density of the fluid, ρ_0 the undisturbed density, and $P(\rho)$ the pressure. The acoustic approximation must be used to derive Eq. (7) from the density equation, the equation of state for the fluid, and Newton's second law of motion. If the fluid is specified to be air, then $\overline{v}$ is the velocity of sound. Of course, for acoustical waves and vibrating strings there may be three-dimensional motion in the general case, satisfying Eq. (1) with the appropriate parameters. *See* SOUND; WAVE MOTION IN FLUIDS.

Electromagnetic waves. A third example of a classical physical situation in which Eq. (1) gives a description of the phenomena is afforded by electromagnetic waves. In a region of space in which the charge and current densities are zero, Maxwell's equations for the photon lead to the wave equations (8). Here $\mathbf{E}$ is

$$\begin{aligned}\left[\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right]\mathbf{E}(\mathbf{r}, t) = 0\\ \left[\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right]\mathbf{B}(\mathbf{r}, t) = 0\end{aligned} \tag{8}$$

the electric field strength and $\mathbf{B}$ is the magnetic flux density; they are both vectors in ordinary space. The parameter c is the speed of light in vacuum. A vector potential $\mathbf{A}(\mathbf{r}, t)$ and a scalar potential $\phi(\mathbf{r}, t)$ may be introduced by the identifications given in Eqs. (9),

$$\begin{aligned}\mathbf{B}(\mathbf{r}, t) &= \nabla \times \mathbf{A}(\mathbf{r}, t)\\ \mathbf{E}(\mathbf{r}, t) &= -\nabla\phi - \frac{1}{c}\frac{\partial \mathbf{A}}{\partial t}\end{aligned} \tag{9}$$

where $\nabla\times$ is the curl operator and ∇ the gradient operator. The potentials then also satisfy the wave equation. *See* POTENTIALS.

If relativistic notation is used so that $A_x = A_1$, $A_y = A_2$, $A_z = A_3$, and $A_4 = i\phi$ with $i = \sqrt{-1}$, then Eq. (10) can be derived (all Greek indices run from

$$\left[\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right]A_\mu(\mathbf{r}, t) = 0 \tag{10}$$

1 to 4). This wave equation may be written in a form that is manifestly covariant (form invariant) with respect to the Lorentz transformations by the further identifications $x = x_1$, $y = x_2$, $z = x_3$, and $x_4 = ict$ so that the operator in Eq. (9) can be written as Eq. (11),

$$\begin{aligned}\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\\ = \frac{\partial^2}{\partial {x_1}^2} + \frac{\partial^2}{\partial {x_2}^2} + \frac{\partial^2}{\partial {x_3}^2} + \frac{\partial^2}{\partial {x_4}^2} \equiv \frac{\partial}{\partial x_\mu}\frac{\partial}{\partial x_\mu}\end{aligned} \tag{11}$$

where the repeated index in the second equation is defined to mean that a sum is to be taken, as indicated above (Einstein summation convention). Then the manifestly covariant form of Eq. (10) is written as Eq. (12). Both $x = (\mathbf{r}, ict)$ and $A = (\mathbf{A}, i\phi)$ are Lorentz

$$\frac{\partial}{\partial x_\mu}\frac{\partial}{\partial x_\mu}A_\nu(\mathbf{r}, t) = 0 \tag{12}$$

four-vectors, and the product of two four-vectors, summed over the repeated index, is a scalar, such as

$$\frac{\partial}{\partial x_\mu}\frac{\partial}{\partial x_\mu}$$

See ELECTROMAGNETIC RADIATION; MAXWELL'S EQUATIONS; RELATIVITY; WAVE (PHYSICS); WAVE MOTION.

Quantum wave equation. The above discussion leads to the consideration of the quantum-mechanical wave equations. The nonrelativistic Schrödinger equation is an example. In this article only the relativistic quantum-mechanical wave equations are considered. *See* NONRELATIVISTIC QUANTUM THEORY.

Schrödinger-Klein-Gordon equation. The first example may be found from Eq. (12) by subtracting from the operator $(\partial/\partial x_\mu)/(\partial/\partial x_\mu)$ a constant term $(m^2c^2)/\hbar^2$, where m is the mass and $\hbar$ is Planck's constant divided by 2π. The new equation is Eq. (13) and is

$$\left[\frac{\partial}{\partial x_\mu}\frac{\partial}{\partial x_\mu} - \frac{m^2c^2}{\hbar^2}\right]\varphi(\mathbf{r}, t) = 0 \tag{13}$$

called the Schrödinger-Klein-Gordon equation. Every free elementary particle has a wave function $\varphi(\mathbf{r}, t)$ which must satisfy this equation although $\varphi(\mathbf{r}, t)$ are different types of functions for particles with different spins. If $\varphi(\mathbf{r}, t)$ has the form of a plane wave given by Eq. (14) corresponding to a particle with momen-

$$\varphi(\mathbf{r}, t) = \chi(\mathbf{P})\exp\left[\frac{i}{\hbar}(\mathbf{P}\cdot\mathbf{x} - Et)\right] \tag{14}$$

tum $\mathbf{P}$ and total energy E, then insertion of Eq. (14) into Eq. (13) yields Eq. (15). $\chi(\mathbf{P})$ in Eqs. (14) and

(15) can be any function of the momentum **P**. Since

$$\frac{1}{\hbar^2}\left[E^2 - c^2p^2 - m^2c^4\right]\chi(\mathbf{P}) = 0 \qquad (15)$$

the term in brackets in Eq. (15) is the relativistic connection between energy and momentum for a free particle, it is identically zero; therefore $\varphi(\mathbf{r}, t)$ defined by Eq. (14) satisfies Eq. (13). Historically, Eq. (15) came before Eq. (13).

Dirac equation. All the above wave equations are second-order in both space and time derivatives. A different type of relativistic quantum-mechanical wave equation is the Dirac equation, which is first-order in both space and time derivatives and has the form shown in Eq. (16). Here α is a three-component

$$\left[\frac{c\hbar}{i}\alpha \cdot \nabla + mc^2\beta\right]\psi(\mathbf{r}, t) = i\hbar\frac{\partial\psi(\mathbf{r}, t)}{\partial t} \qquad (16)$$

object whose components are each a 4×4 matrix, β is a 4×4 matrix, and the 4×1 column matrix $\psi(\mathbf{r}, t)$ is a wave function corresponding to a particle of intrinsic spin $^1/_2$.

If ψ_1, ψ_2, ψ_3, ψ_4 are the four components of $\psi(\mathbf{r}, t)$, then Eq. (16) is the matrix form of the system of Eq. (17), where $\alpha_i^{\sigma\tau}$ is the $\sigma\tau$ element of

$$\frac{c\hbar}{i}\sum_{\tau=1}^{4}\left(\alpha_1^{\sigma\tau}\frac{\partial}{\partial x_1} + \alpha_2^{\sigma\tau}\frac{\partial}{\partial x_2} + \alpha_3^{\sigma\tau}\frac{\partial}{\partial x_3}\right)\psi_\tau + \sum_{\tau=1}^{4}\beta_{\sigma\tau}mc^2\Psi_\tau = i\hbar\frac{\partial\psi_\sigma}{\partial t} \qquad (17)$$

the 4×4 matrix which forms the ith component of α in Eq. (16). Because it is necessary that each component of the solution $\psi(\mathbf{r}, t)$ of Eq. (16) also satisfy Eq. (13), α and β have certain anticommutation relations with each other which imply that the product of any of these matrices with itself is the 4×4 unit matrix. Equation (16) may be written in a manifestly covariant form by making the definitions $\gamma \equiv -i\beta\alpha$, $\gamma_4 \equiv \beta$ and by rearranging terms in Eq. (16). The result is (in index notation) Eq. (18).

$$\left[\gamma_\mu\frac{\partial}{\partial x_\mu} + \frac{mc}{\hbar}\right]\psi(\mathbf{r}, t) = 0 \qquad (18)$$

The term $\beta\alpha$ is interpreted as a three-component object with matrices $\beta\alpha_1$, $\beta\alpha_2$, and $\beta\alpha_3$ as components. The quantum-mechanical functions such as $\psi(\mathbf{r}, t)$ and $\varphi(\mathbf{r}, t)$ differ from their classical counterparts $y(\mathbf{r}, t)$ and so forth, in that the quantum functions are not directly observable quantities and are, in fact, complex. In general, only bilinear combinations of the quantum-mechanical wave functions, suitably averaged with certain operators, and over some region of space, are physically observable.

The Dirac equation is the most prominent example of the type of wave equation (relativistic quantum mechanical) that is linear in the space and time derivatives.

Rarita-Schwinger equation. A second example is the Rarita-Schwinger equation (19). This has the same

$$\left[\gamma_\mu\frac{\partial}{\partial x_\mu} + \frac{mc}{\hbar}\right]\psi_\nu(\mathbf{r}, t) = 0 \qquad (19)$$

form as the Dirac equation except that the wave function $\psi_\nu(\mathbf{r}, t)$ is now a 4×1 column matrix with each of the four components being independently a four-vector so that $\psi_\nu(\mathbf{r}, t)$ has 16 components. Equation (19) is intended to describe a free relativistic particle (and antiparticle) with intrinsic spin $^3/_2$. For this, only eight independent complex components are needed; in addition to Eq. (19), the auxiliary condition defined by Eq. (20) is required.

$$\gamma_\nu\psi_\nu(\mathbf{r}, t) = 0 \qquad (20)$$

Equations (19) and (20) together imply a third equation, Eq. (21). Exactly as in the Dirac equation,

$$\frac{\partial}{\partial x_\nu}\psi_\nu(\mathbf{r}, t) = 0 \qquad (21)$$

each component of the Rarita-Schwinger wave function satisfies Eq. (13). The necessity of auxiliary conditions, such as Eqs. (20) and (21), to eliminate redundant components in the wave function is a characteristic of the approach to relativistic wave equations which involve only first derivatives of space and time and certain matrices. These considerations for spins $^1/_2$ and $^3/_2$ may be extended to higher half-integral spins by including more four-vector indices on the wave function.

Duffin-Kemmer-Petiau equation. Another way of writing such a first-order matrix-differential wave equation is to make all the components of the wave function matrix components. The wave function $\overline{\psi}(\mathbf{r}, t)$ is then a matrix satisfying a wave equation of the form of Eq. (22), where the four β_μ are a set of singular matrices whose structure depends on the intrinsic spin of the particle to be described. For spin zero the β_μ are 5×5 matrices and for spin 1 they are 10×10. Of course, each component of $\overline{\psi}(\mathbf{r}, t)$ must also satisfy Eq. (13). Equation (22) is called the Duffin-Kemmer-Petiau equation.

$$\left[\beta_\mu\frac{\partial}{\partial x_\mu} + \frac{mc}{\hbar}\right]\overline{\psi}(\mathbf{r}, t) = 0 \qquad (22)$$

Tensor equations. Free, relativistic quantum-mechanical particles with tensor equations may also be described. The best-known examples of this type are the Proca equations for spin 1 which have the form shown in Eqs. (23), and so are differential relations

$$F_{\alpha\beta} = \frac{\partial\omega_\beta}{\partial x_\alpha} - \frac{\partial\omega_\alpha}{\partial x_\beta} \qquad (23a)$$

$$\frac{\partial F_{\alpha\beta}}{\partial x_a} = \frac{m^2c^2}{\hbar^2}\omega_\beta \qquad (23b)$$

between a four-vector field $\omega_\beta(\mathbf{r}, t)$ and a second-rank tensor field $F_{\alpha\beta}(\mathbf{r}, t)$. By differentiating Eq. (23*b*) with respect to x_β, the auxiliary condition shown in Eq. (24) can be obtained; by inserting Eq. (23*a*) into

$$\frac{\partial}{\partial x_\beta}\omega_\beta(\mathbf{r}, t) = 0 \qquad (24)$$

Eq. (23*b*) and using Eq. (24), Eq. (13) can be satisfied by each of the components of ω_β. Thus it is again apparent that auxiliary conditions are needed to

eliminate redundant components from the quantum-mechanical equations of motion.

Spinor equations. The most general set of wave equations and auxiliary conditions capable of describing quantum-mechanically a free relativistic particle with mass and internal integral or half-integral spin are the Dirac-Fierz-Pauli spinor wave equations.

Equations with minimum number of components. A characteristic of all the above quantum-mechanical formulations is that the function satisfying the wave equation has more components than necessary to describe a free particle and antiparticle with the given spin. A description receiving much attention is that involving only irreducible representations of the homogeneous Lorentz group.

With this description a system with spin s is described by a wave function with $2s + 1$ components for the particle and $2s + 1$ components for the antiparticle, the minimum number for describing such a system. For the wave equation in this description, there may be only first-order time derivatives and complicated spatial derivatives, or there may be uniform dependence of the wave equation on both space and time derivatives, the order depending on the spin.

In many cases involving the $2(2s + 1)$-component description of a particle-antiparticle with intrinsic spin s, it is possible to decouple the two sets of $2s + 1$ components in the wave function by setting the mass m to zero. The wave equation then applies to a massless particle or antiparticle. For example, Eq. (16) with $m = 0$ becomes Eq. (25), in which

$$\frac{c\hbar}{i}\alpha \cdot \nabla\psi(\mathbf{r}, t) = i\hbar\frac{\partial\psi}{\partial_t}(\mathbf{r}, t) \tag{25}$$

the wave function $\psi(\mathbf{r}, t)$ now also refers to $m = 0$. With the particular choice of representation of Dirac matrices of Eq. (26), where σ are the Pauli spin ma-

$$\alpha = \begin{pmatrix} \sigma & 0 \\ 0 - & \sigma \end{pmatrix} \tag{26}$$

trices, the upper two components of the wave function are decoupled from the lower two components. The wave equation for the latter is Eq. (27), where

$$-\frac{c\hbar}{i}\sigma \cdot \nabla\chi(\mathbf{r}, t) = i\hbar\frac{\partial}{\partial_t}\chi(\mathbf{r}, t) \tag{27}$$

$\chi(\mathbf{r}, t)$ is the lower two components of $\psi(\mathbf{r}, t)$. Equation (27) is known as the two-component neutrino equation, and the related equation for the upper two components represents a wave equation for the antineutrino, the antiparticle of a neutrino.

Wave equations like Eq. (27), with the Pauli matrices replaced by $\mathbf{s}/s$ and $\chi(\mathbf{r}, t)$ having $2s + 1$ components, are known to be appropriate for spin s massless particles when auxiliary conditions are included to restrict the wave function to two independent components as obtained in the two-component neutrino theory. This formulaton also includes the photon as a spin-1 example of a massless particle and offers an alternative description to Maxwell's equations discussed above. Such a description shows great promise for future developments in the field of relativistic quantum-mechanical wave equations. *See* QUANTUM MECHANICS; RELATIVISTIC QUANTUM THEORY.

David L. Weaver

Bibliography. J. D. Bjorken and S. D. Drell, *Relativistic Quantum Mechanics*, 1964; E. M. Corson, *Introduction to Tensors, Spinors and Relativistic Wave Equations*, 2d ed., 1981; C. A. Coulson and A. Jeffrey, *Waves: A Mathematical Approach to the Common Types of Wave Motion*, 2d ed., 1978; C. Itzykson and J. B. Zuber, *Quantum Field Theory*, 1980; G. A. Korn and T. M. Korn, *Mathematical Handbook for Scientists and Engineers*, 2d ed., 1968; P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, 2 vols., 1953; M. D. Scadron, *Advanced Quantum Theory and Its Application through Feynman Diagrams*, 2d ed., 1991.

Wave mechanics

The modern theory of matter holding that elementary particles (such as electrons, protons, and neutrons) have wavelike properties. By 1915, experiments on the diffraction (bending) of x-rays into special directions by crystals had established that x-rays were electromagnetic waves, akin to visible and infrared light, but of much shorter wavelength than those other electromagnetic radiations. However, in 1923 A. H. Compton showed that observations on x-rays scattered by, for example, a graphite target could be quantitatively predicted via the hypothesis that the x-ray scattering from each individual target atom resulted from elastic (billiard-ball-like) collisions between the comparatively slowly moving atomic electrons and what were in effect particles (now commonly called photons) in the incident x-radiation, with each incident photon having an energy equal to the product of Planck's constant (6.6×10^{-34} joule second) and the speed of light divided by the wavelength, and momentum equal to Planck's constant divided by the wavelength. Thus, some experiments with electromagnetic radiation seemingly can be understood only by visualizing the radiation as waves, while other experiments on the same radiation seemingly require that the radiation be visualized as a stream of particles. This quite unintuitive wave-particle duality manifested by electromagnetic radiation is all the more remarkable in that the particlelike properties associated with the radiation, namely the energy and momentum of its photons, are given in terms of its wavelength, a concept that seemingly has no meaning except in a wave context. *See* COMPTON EFFECT; PHOTON; X-RAY DIFFRACTION.

In 1924 L. de Broglie postulated that the wave-particle duality which had been demonstrated for electromagnetic radiation also was a property of the elementary particles, such as electrons and protons, making up the atoms and molecules forming ordinary matter. In particular, de Broglie postulated that a particle has an associated wavelength obeying the same relation as was found to hold for photons, namely: the wavelength equals Planck's

constant divided by the particle's momentum (as customarily defined in elementary mechanics). This hypothesis was verified in 1927 in an experiment in which a beam of electrons having known momentum is diffracted by a crystal into special directions. As in the x-ray diffraction experiments discussed above, such diffraction seems understandable only on the hypothesis that the electrons are waves. Furthermore, the wavelength of the electrons in the incident beam, computed via the very same formula as was used to derive the x-ray wavelengths in the x-ray diffraction experiments, agreed precisely with the de Broglie relation. *See* ELECTRON DIFFRACTION.

Subsequent experiments have confirmed that not merely electrons but material particles in general, such as neutrons and neutral sodium atoms, manifest the wave-particle duality and obey the de Broglie relation. The de Broglie relation and the qualitative wave-particle duality concept have been incorporated into the highly successful modern theory of quantum mechanics. *See* DE BROGLIE WAVELENGTH; INTERFERENCE OF WAVES; QUANTUM MECHANICS. Edward Gerjuoy

Bibliography. E. Fermi and L. Marshall, Interference phenomena of slow neutrons, *Phys. Rev.*, 71:666–667, 1947; D. Halliday, R. Resnick, and K. S. Krane, *Physics*, vol. 2, 4th ed., 1992; D. W. Keith et al., Diffraction of atoms by a transmission grating, *Phys. Rev. Lett.*, 61:1580–1583, 1988; F. K. Richtmyer, E. H. Kennard, and J. N. Cooper, *Introduction to Modern Physics*, 6th ed., 1969.

Wave motion

The process by which a disturbance at one point in space is propagated to another point more remote from the source with no net transport of the material of the medium itself. For example, sound is a form of wave motion; wind is not. Wave motion can occur only in a medium in which energy can be stored in both kinetic and potential form. In a mechanical medium, kinetic energy results from inertia and is stored in the velocity of the molecules, while potential energy results from elasticity and is stored in the displacement of the molecules.

In a free traveling wave (as distinguished from a stationary or standing wave) one part of the medium disturbs an adjacent part, thereby imparting energy to it. This portion of the medium, in turn, disturbs another part, thereby causing a flow of energy in a given direction away from the source (see **illus.**). More technically, wave propagation is the result of kinetic energy at one point being transferred into potential energy at an adjacent point, and vice versa. The rate of travel of the disturbance, or velocity of propagation, is determined by the constants of the medium. A stationary wave is the combination of two waves of the same frequency and strength traveling in opposite directions so that no net transfer of energy away from the source takes place. A standing wave is the same but with the returning wave (toward the source) being of lesser intensity than the outwardly traveling wave so that a net transfer of energy away from the source does take place.

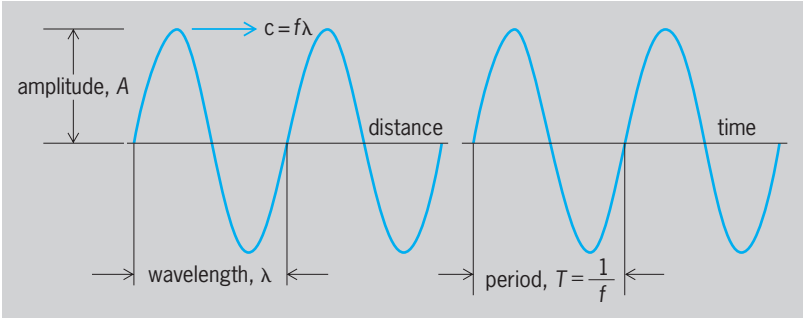
Wave motion can occur in a vacuum (electromagnetic waves), in gases (sound waves), in liquids (hydrodynamic waves), and in solids (vibration waves). Electromagnetic waves can also travel in gases, liquids, and solids provided that the electrical conductivity of the medium is not perfect or that the imaginary part of the dielectric constant is not infinitely great. By current usage, elastic waves propagated in gases, liquids, and solids, regardless of whether one can hear them or not, are called acoustic waves.

In this article, it is shown that wave motion can exist in any continuous medium in which potential and kinetic energy can be stored. The wave equation is derived, and solutions for electromagnetic and mechanical media are given. It is shown that longitudinal waves may exist in all mechanical media, and that transverse waves may exist in electromagnetic and solid mechanical media. In all cases, viscosity or heat losses in mechanical media, and resistance and magnetic losses in electromagnetic media are neglected. Only homogeneous and isotropic media are considered. For information related to the present discussion *see* DIFFRACTION; ELECTROMAGNETIC RADIATION; HARMONIC MOTION; HUYGENS' PRINCIPLE; INTERFERENCE OF WAVES; LIGHT; MAXWELL'S EQUATIONS; REFRACTION OF WAVES; SOUND; SUPERPOSITION PRINCIPLE; VIBRATION; WAVE EQUATION; WAVE MOTION IN LIQUIDS.

Fundamental relations. A wave is commonly referred to in terms of either its wavelength or its frequency. In any type of wave motion, these two quantities are related to a third quantity, velocity of propagation, by Eq. (1), where f = frequency, λ =

$$f\lambda = c \tag{1}$$

wavelength, and c = velocity of propagation. The period T is the reciprocal of the frequency, and the amplitude A is the maximum magnitude taken on by the variable of the wave at a given point in space. The physical significance of these quantities is presented in the illustration. It is a basic property of wave motion that the frequency of a wave remains constant under all circumstances except for a relative motion between the source of the wave and the



Relation between frequency, wavelength, and velocity in wave propagation.

observer. The case of a frequency shift due to relative motion, known as the Doppler effect, has many interesting applications. On the other hand, the velocity of propagation is dependent on the properties of the medium (and sometimes also on the frequency), and the wavelength will vary with the velocity in accordance with Eq. (1). *See* DOPPLER EFFECT.

Electromagnetic waves. The media in which electromagnetic waves travel possess no elasticity or inertia, but the ability to store energy in the electric and magnetic fields. The electric field corresponds in every respect to the field of an irrotational fluid motion, and the mathematical formulations of the motions of acoustic waves and electromagnetic waves are similar.

In the customary manner, in an electromagnetic medium one defines the vectorial factors **E** as the electric field strength and **H** as the magnetic field strength, with both magnitudes and directions at every point in space.

Next, it is assumed that the medium is homogeneous and isotropic (that is, that the dielectric constant κ and the permeability μ are constants and scalars), that there are no applied electromotive forces in the portion of the medium being dealt with, and that the electrical conductivity σ of the medium is zero. (Waves can be propagated in media where σ is greater than zero but less than infinity.) Then the field equations can be written as Eqs. (2)–(5), where

$$\frac{\kappa}{c}\frac{\partial \mathbf{E}}{\partial t} = \text{curl } \mathbf{H} \tag{2}$$

$$-\frac{\mu}{c}\frac{\partial \mathbf{H}}{\partial t} = \text{curl } \mathbf{E} \tag{3}$$

$$\text{div } \mathbf{H} = 0 \tag{4}$$

$$\text{div } \mathbf{E} = 0 \tag{5}$$

c will be shown to be the velocity of propagation of the electromagnetic wave in a vacuum, where $\kappa = \mu = 1$. *See* CALCULUS OF VECTORS.

Wave equation. J. C. Maxwell recognized about 1863 that these basic equations could be combined to yield an equation resembling the wave equation for mechanical wave motion. Thus he predicted the existence of electromagnetic waves which had not been suspected before. Later, electromagnetic waves proved to be identical with light waves.

The combination of Eqs. (2)–(5) yields the wave equation (6). In a vacuum, the wave equation becomes Eq. (7) and also Eq. (8), where ∇^2 is a scalar

$$\frac{\kappa\mu}{c^2}\frac{\partial^2 \mathbf{E}}{\partial t^2} = \nabla^2 \mathbf{E} \tag{6}$$

$$\frac{\partial^2 \mathbf{E}}{\partial t^2} = c^2 \nabla^2 \mathbf{E} \tag{7}$$

$$\frac{\partial^2 \mathbf{H}}{\partial t^2} = c^2 \nabla^2 \mathbf{H} \tag{8}$$

operator called the laplacian. *See* LAPLACIAN.

Plane wave propagation. For simplicity, particular solutions of the wave equation for the case of plane wave propagation will now be sought. A wave is called plane when a family of parallel planes can be taken in the field such that the electric and magnetic field strengths are constant in magnitude and direction at all points of any given member of the family; the planes are called wavefronts, the direction perpendicular to them the wave normal. If the axis of x is taken in the direction of the wave normal, then the wavefronts are parallel to the plane of yz.

Since, for this case, **E** and **H** are to be constant in any one wavefront, the partial derivatives with respect to y or z must vanish. Then the field equations reveal that Eq. (9) is valid. Therefore, insofar as wave

$$\frac{\partial E_x}{\partial t} = \frac{\partial E_x}{\partial x} = \frac{\partial H_x}{\partial t} = \frac{\partial H_x}{\partial x} = 0 \tag{9}$$

propagation is concerned, Eq. (10) holds.

$$E_x = H_x = 0 \tag{10}$$

The meaning of Eq. (10) is that neither **E** nor **H** can have a periodically changing component in the direction in which the wave is traveling. That is to say, the waves are not longitudinal but transverse.

Of the remaining four equations developing out of the field equations, two of them connect E_y and H_z, and the other two E_z and H_y. Dealing with the pairs, one obtains the one-dimensional wave equation (11),

$$\frac{\partial^2(\,)}{\partial t^2} = c^2\frac{\partial^2(\,)}{\partial x^2} \tag{11}$$

where E_y, E_z, H_y, or H_z may be inserted within the parentheses.

The general solution of Eq. (11) can be written in the form of Eq. (12). is The first term of Eq. (12)

$$E_y = F_1(x - ct) + F_2(x + ct) \tag{12}$$

associated with a wave traveling in the positive x direction, and the second term with a wave traveling in the minus x direction. This is true because the argument of F_1 is the same whenever $x = ct$, so that as t becomes greater, the position of a wavefront moves in the positive direction of x. Since for F_2, $-x = ct$, the opposite is true.

If only the outward traveling wave is considered, Eq. (13) holds. Such a wave is transmitted without change of shape, because its position x can always

$$E_y = F_1(x - ct) \tag{13}$$

equal ct. That is, at any given elapsed time t, E_y will have the same value at $x = ct$ as it had at $x = t = 0$. Also, because $c = x/t$, c has the dimensions of a velocity and is the speed at which the wave travels.

Light is also an electromagnetic process so that one should be able to state the optical properties of a substance once its electrical constants are given. However, at the high frequencies of visible light, there is generally dispersion in the medium so that

κ and μ vary with frequency, and this variation must be taken into account if Maxwell's equations are to give a description of optical phenomena which fits measured data.

Acoustic waves in gases. In gases, the existence of inertia (resulting in kinetic energy) and elasticity (resulting in potential energy) is obvious. For example, imagine taking from the medium a small packet of gas enclosed by a weightless deformable membrane. When this packet is squeezed, it is found to have stiffness (or elasticity). Its mass equals the mass of the air molecules contained therein. Also, the deformable membrane keeps the mass inside constant.

From this description, it can be seen that it is a relatively simple matter to describe the motion of such a packet as a function of time. The mass element is labeled in any convenient manner, the most common being by its location at any convenient time. This method is called the lagrangian description of motion.

To derive the equations for wave motion in a gas, one expresses the concepts of inertia by Newton's second law, of elasticity by the perfect gas law, and of the deformable packet by the conservation of mass law. The packet is assumed to be rectangular in shape. To find the equation of motion, the supposition is made that the box is situated in a medium in which the pressure p changes in space at a space rate given by Eq. (14), where **i**, **j**, and **k** are unit vectors

$$\text{grad } p \equiv \mathbf{i}\frac{\partial p}{\partial x} + \mathbf{j}\frac{\partial p}{\partial y} + \mathbf{k}\frac{\partial p}{\partial z} \tag{14}$$

in the x, y, and z directions, respectively, and p is the pressure at a point.

The net force **f** acting to move the box in some direction is equal to the vector summation of the gradients in force across the three pairs of faces of the packet times the respective separations of these faces; in the positive direction, Eq. (15) holds. Here

$$\mathbf{f} = -\left[\mathbf{i}\left(\frac{\partial p}{\partial x}\Delta x\right)\Delta y\Delta z + \mathbf{j}\left(\frac{\partial p}{\partial y}\Delta y\right)\Delta x\Delta z + \mathbf{k}\left(\frac{\partial p}{\partial z}\Delta z\right)\Delta x\Delta y\right] = -V\text{grad } p \tag{15}$$

V is the the average volume of the packet. The positive gradient causes an acceleration of the box in the negative direction of x.

By Newton's second law, the force acting to accelerate the packet is given by Eq. (16), where **q** is the

$$\mathbf{f} = \rho' V\frac{\partial \mathbf{q}}{\partial t} \tag{16}$$

average vector velocity of the gas in the packet, ρ' the average density of the gas in the packet, and $\rho' V$ the total (constant) mass of the gas in the packet. In writing Eq. (16), it has been assumed that the packet is never displaced from its equilibrium position by a significant part of a wavelength of sound.

Combining Eqs. (15) and (16) yields Eq. (17). In keeping with the approximation just stated, it has been assumed that ρ' (the instantaneous density) does not appreciably deviate from the average density ρ_0. These approximations are acceptable, providing the sound pressures are below about 100 dynes/cm^2 (10 newtons/m^2).

$$-\text{grad } p = \rho_0\frac{\partial \mathbf{q}}{\partial t} \tag{17}$$

Assuming adiabatic expansions and contractions of the gas, and that the gas is perfect, Eq. (18) is

$$\frac{dP}{P} = -\frac{\gamma\, dV}{V} \tag{18}$$

obtained. Here P is the total pressure in the gas and γ is the ratio of specific heat at constant pressure to specific heat at constant volume.

P and V can be defined by Eqs. (19), where p and

$$\begin{aligned} P &= P_0 + \\ V &= V_0 + \tau \end{aligned} \tag{19}$$

τ are time-varying quantities, and P_0 and V_0 are equilibrium values. If inequalities (20) hold, Eq. (21) is valid.

$$\begin{aligned} p &\ll P_0 \\ \tau &\ll V_0 \end{aligned} \tag{20}$$

$$\frac{1}{P_0}\frac{\partial p}{\partial t} = \frac{-\gamma}{V_0}\frac{\partial \tau}{\partial t} \tag{21}$$

To satisfy the law of mass conservation, one writes Eq. (22) where ξ is the average vector displacement

$$\tau = V_0 \text{ div } \xi \tag{22}$$

of the box.

Differentiation of Eq. (22) with respect to time, and substitution of it in Eq. (21), yields Eq. (23).

$$\frac{\partial p}{\partial t} = -\gamma P_0 \text{ div } \mathbf{q} \tag{23}$$

The elimination of p from Eqs. (17) and (23) yields the wave equation (24), where by definition Eq. (25) applies.

$$\frac{\partial^2 p}{\partial t^2} = c^2\nabla^2 p \tag{24}$$

$$c^2 = \frac{\gamma P_0}{p_0} \tag{25}$$

One-dimensional plane waves. An acoustic wave is called plane when a family of parallel planes can be taken in the medium such that the pressures and particle velocities are constant in magnitude and the particle velocities are constant in direction at all points of any given member of the family.

Since p and **q** are to be constant in any one wavefront, the partial derivatives with respect to y or z must vanish, and in Eq. (23), Eq. (26) applies. Since

$$\text{div } \mathbf{q} = \frac{\partial q_x}{\partial x} \tag{26}$$

Eq. (26) reveals that the x component of **q** is the only component of **q** remaining in the equations basic to the wave equation, the wave is longitudinal.

The one-dimensional wave equation becomes Eq. (27).

$$\frac{\partial^2 p}{\partial t^2} = c^2 \frac{\partial^2 p}{\partial x^2} \tag{27}$$

The general solution to Eq. (27) is exactly the same as that for electromagnetic waves and is given by Eq. (12). The discussion following Eq. (12) is also valid here.

One-dimensional spherical waves. In free space, it is frequently desired to express mathematically the radiation of sound from a nondirectional source. In this case the sound wave expands as it travels away from the source and the wavefront is always a spherical surface. The operator on the right side of Eq. (24) can be written in a form suitable to spherical coordinates.

Assuming equal radiation in all directions, the wave equation in one-dimensional spherical coordinates then becomes Eq. (28).

$$\frac{\partial^2 p}{\partial r^2} + \frac{2}{r}\frac{\partial p}{\partial r} = \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} \tag{28}$$

Differentiation shows that Eq. (28) can also be written as Eq. (29).

$$\frac{\partial^2 (pr)}{\partial t^2} = c^2 \frac{\partial^2 (pr)}{\partial r^2} \tag{29}$$

Equation (29) has the same form as that of Eq. (27). Hence, the same formal solution applies to either equation, except that the dependent variable is $p(x,t)$ in one case and $pr(r, t)$ in the other case.

The solution to Eq. (29) for the outward traveling wave only (free space) is thus given by Eq. (30).

$$p = \frac{1}{r} F_1(r - ct) \tag{30}$$

Note that, just as for the plane wave, the wave is propagated without change of shape. However, the magnitude of the sound pressure decreases inversely with distance owing to the spreading of the wave as it propagates.

Acoustic waves in liquids. Acoustic waves in liquids obey the same equations as those in gases. The velocity of propagation is greater in liquids than in gases, and viscous losses are often higher. For information on an important example of wave motion in liquids *see* UNDERWATER SOUND.

Elastic and flexural waves in solids. Several different types of acoustic waves may exist in solids, depending on the different manners in which potential energy is stored in the solid. The wave equations associated with several of these types are reviewed in the following paragraphs.

Waves on flexible stretched strings. The wave equation for a flexible stretched string is given by Eq. (31),

$$\frac{\partial^2 \xi_y}{\partial t^2} = c^2 \frac{\partial^2 \xi_y}{\partial x^2} \tag{31}$$

where ξ_y is the displacement of the string at a point x along the string in the y direction (perpendicular to the string). The speed of propagation c is equal to the square root of the ratio of the tension (in dynes) to the linear density of the string (in g/cm^2).

The solution to this equation is identical to that for Eq. (11). Because the motion of the elements of the string is perpendicular to the string, at least for small displacements, the waves are said to be transverse.

Flexural waves in bars. For the purposes of this article, it is assumed that a string has tension with negligible stiffness, while a bar has stiffness without tension. When a bar is bent, its lower half is compressed and its upper half is stretched, or vice versa. When the bending force is removed, the bar attempts to regain its equilibrium position. The restoring force is due to the moment of the forces about the neutral plane in the bar and is related to the cross-sectional dimensions and the Young's modulus of the material.

The transverse wave equation describing the motion of such a bar is Eq. (32), where ξ_y is the displacement perpendicular to the neutral plane of the bar, Y is Young's modulus, ρ is the density of the bar, and κ is the radius of gyration of the cross section. Values of κ for some of the simpler cross-sectional shapes are given by Eqs. (33).

$$\frac{\partial^2 \xi_y}{\partial t^2} = -\kappa^2 \frac{Y}{\rho} \frac{\partial^4 \xi_y}{\partial x^4} \tag{32}$$

$$\kappa = \frac{a}{\sqrt{12}} \tag{33a}$$

Rectangle: Length b parallel to center line, width a perpendicular to center line, Circle: Radius a,

$$\kappa = \frac{a}{2} \tag{33b}$$

Circular ring: Outer radius a, inner radius b,

$$\kappa = 0.5(a^2 + b^2)^{1/2} \tag{33c}$$

Equation (32) differs from the usual wave equation (11) in that it has a fourth derivative with respect to x instead of a second derivative.

The function $F_1(x - ct)$ is not a solution, so that a bar satisfying Eq. (32) cannot have waves traveling along it with a velocity independent of frequency and with an unchanged shape.

If one considers excitation of the bar by one frequency f at a time, a solution to Eq. (32) may be written as Eq. (34), where $\mu = (f^2\rho/4\pi^2 Y\kappa^2)^{1/4}$,

$$\xi_y = A \cos\left[2\pi(\mu x - ft) + \phi\right] \tag{34}$$

$f = 2\pi\mu^2\kappa\sqrt{Y/\rho}$, A = amplitude, and ϕ = phase angle.

The velocity of propagation of the wave is equal to $(f/\mu) = (4\pi^2 Y\kappa^2/\rho)^{1/4}\sqrt{f}$. It obviously depends on the frequency f of the exciting wave. Such a velocity for a simple harmonic wave is called the phase velocity. For a complex wave with several components differing in frequency, the shape must change with distance of travel. A bar is sometimes said to be a dispersing medium for waves of bending. *See* PHASE VELOCITY.

There is also a possibility for a longitudinal wave in such a bar. In this case the bar is excited at an end

in the direction of its longitudinal axis. The wave equation is identical to Eq. (31), except that ξ_x replaces ξ_y. The motions of the particles of the bar are in a line with the longitudinal axis of the bar. Longitudinal and transverse waves may, and generally do, coexist.

Waves in plates. Wave motion in a plate is similar to that in a bar. The bending of a plate compresses the material on the inside of the bend and stretches it on the outside. But when a material is compressed, it tries to spread out in a direction perpendicular to the compressed force, so that when a plate is bent downward in one direction there is a tendency for it to curl up in a direction at right angles to the bend. The ratio of the sideways spreading to the compression is called Poisson's ratio and is designated by the letter σ.

The wave equation for the plate is given by Eq. (35), where η is the displacement of the plate

$$\Delta^4\eta + \frac{12}{(hc_L)^2}\frac{\partial^2\eta}{\partial t^2} = 0 \tag{35}$$

perpendicular to its surface, h is the thickness of the plate, and c_L is the longitudinal plate velocity given by Eq. (36).

$$c_L = \sqrt{\frac{Y}{\rho(1-\sigma^2)}} \tag{36}$$

From a solution to the equation it is possible to find the velocity of propagation for the transverse (bending) wave, as in Eq. (37). Just as for the bar, the

$$c_B = \sqrt{1.8hc_L}\sqrt{f} \tag{37}$$

velocity of propagation is dependent on frequency, and the plate is said to be a dispersing medium.

Steady-state wave motion. Very frequently the source of a wave is steady and the wave produced by it is periodic, at least for a long period of time compared to the buildup and decay times of the wave. When this is true, it is convenient to specify the functions $F_1(x - ct)$ and $F_2(x + ct)$ of Eq. (12) by a summation of sinusoidal functions, as in Eq. (38), where

$$F(x - ct) = \sum_\nu A_\nu \cos\left[\omega_\nu\left(t - \frac{x}{c}\right) + \theta_\nu\right] \tag{38}$$

$\omega_\nu = 2\pi f_\nu$; f_ν is the frequency of the νth component of the wave, and θ_ν is the phase angle of the νth component, which is determined when $x = t = 0$.

From the well-known theory of Fourier series, any function $F_1(x - ct)$, if it repeats on itself periodically, can be represented by a linear summation of sine or cosine wave functions. The component with the lowest frequency f_1 is the fundamental component or the first harmonic; each higher component, called the $(\nu - 1)$th overtone, or the νth harmonic, has a frequency equal to νf_1. In other words, ν is an integer, and the waveform $F_1(x - ct)$ is represented by a linear summation of cosine terms with frequency components f_ν that are harmonically related (f, $2f$, $3f$, $4f$, and so forth) and at $t = x = 0$ have phase angles $\theta_1, \theta_2, \theta_3$, and so forth, that may differ from zero. *See* FOURIER SERIES AND TRANSFORMS.

As a further simplification, each component of Eq. (38) is usually written as Eq. (39), where $k = \omega/c$

$$P = A\cos(\omega t - kx + \theta) \tag{39}$$

is called the wave number. The period of this component is T and equals $1/f$. As before, θ is the phase angle in radians.

Note that A is the peak amplitude of the component wave. Generally, in acoustics, measuring devices read the root-mean-square value of a wave, so that the intensity of the wave is designated by $A_{\text{rms}} = A/\sqrt{2}$, and the strength of a wave represented by Eq. (38) is given by Eq. (40).

$$A_{\text{rms}} = \sqrt{A_{1_{\text{rms}}}^2 + A_{2_{\text{rms}}}^2 + \cdots} \tag{40}$$

See ROOT-MEAN-SQUARE.

Leo L. Beranek

Bibliography. L. L. Beranek, *Acoustics*, 1954, paper 1986; H. Georgi, *The Physics of Waves*, 1992; L. E. Kinsler et al., *Fundamentals of Acoustics*, 4th ed., 1999; P. Lorrain, *Electromagnetic Fields and Waves*, 3d ed., 1995; I. G. Main, *Vibrations and Waves in Physics*, 3d ed., 1993; H. J. Pain, *Physics of Vibrations and Waves*, 5th ed., 1999.

Wave optics

The branch of optics which treats of light (or electromagnetic radiation in general) with explicit recognition of its wave nature. The counterpart to wave optics is ray optics or geometrical optics, which does not assume any wave character but treats the propagation of light as a straight-line phenomenon except for changes of direction induced by reflection or refraction. *See* GEOMETRICAL OPTICS; OPTICS.

Any optical phenomenon which is correctly describable in terms of geometrical optics can also be correctly described in terms of wave optics. However, the many phenomena of interference, diffraction, and polarization are incontrovertible evidence of the wave nature of light, and geometrical optics often gives an incomplete or incorrect description of the behavior of light in an optical system. This is especially true if changes of refractive index occur within a space which is of the order of several wavelengths of the light. *See* DIFFRACTION; ELECTROMAGNETIC WAVE; INTERFERENCE OF WAVES; POLARIZED LIGHT.

Richard C. Lord

Wave-shaping circuits

Electronic circuits used to create or modify specified time-varying electrical voltage or current waveforms using combinations of active electronic devices, such as transistors or analog or digital integrated circuits, and resistors, capacitors, and inductors. Most wave-shaping circuits are used to generate periodic waveforms. *See* INTEGRATED CIRCUITS; TRANSISTOR.

The common periodic waveforms include the square wave, the sine and rectified sine waves, the sawtooth and triangular waves, and the periodic arbitrary wave. The arbitrary wave can be made to conform to any shape during the duration of one period. This shape then is followed for each successive cycle. *See* FUNCTION GENERATOR; WAVEFORM.

A number of traditional electronic and electromechanical circuits are used to generate these waveforms. Sine-wave generators and *LC*, *RC*, and beat-frequency oscillators are used to generate sine waves; rectifiers, consisting of diode combinations interposed between sine-wave sources and resistive loads, produce rectified sine waves; multivibrators can generate square waves; electronic integrating circuits operating on square waves create triangular waves; and electronic relaxation oscillators can produce sawtooth waves. *See* ALTERNATING CURRENT; DIODE; MULTIVIBRATOR; OPERATIONAL AMPLIFIER; OSCILLATOR; RECTIFIER.

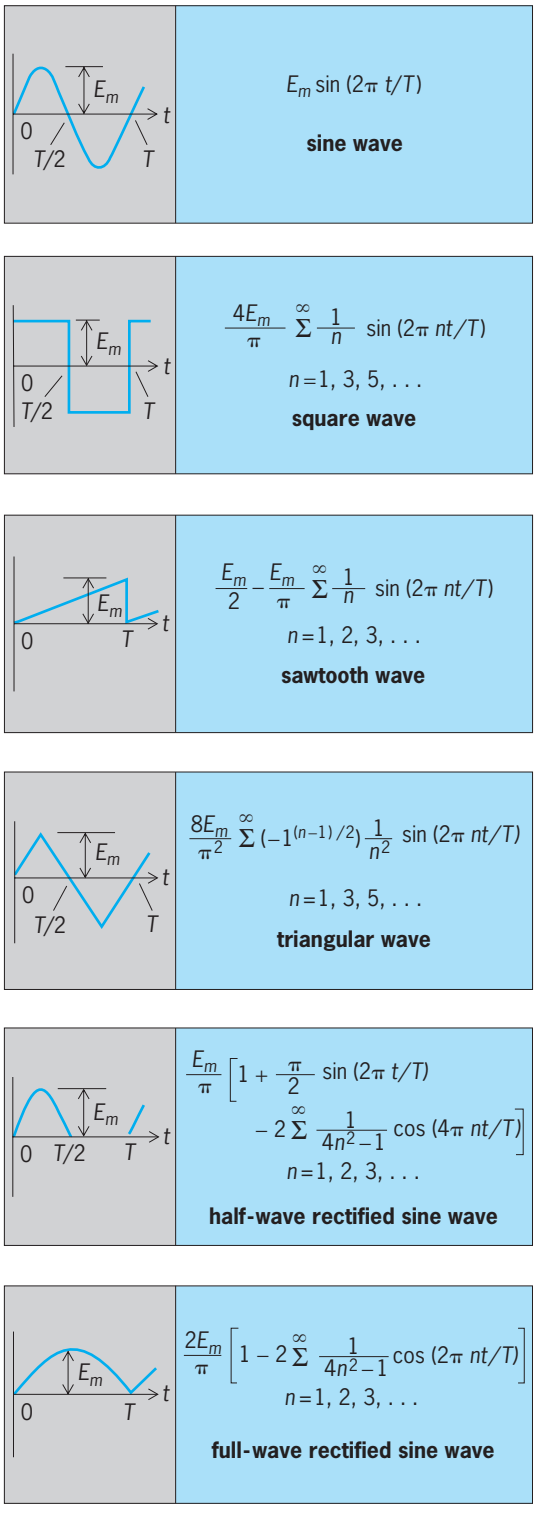
In many applications, generation of these standard waveforms is now implemented using digital circuits. Digital logic or microprocessors generate a sequence of numbers which represent the desired waveform mathematically. These numerical values then are converted to continuous-time waveforms by passing them through a digital-to-analog converter. Digital waveform generation methods have the ability to generate waveforms of arbitrary shape, a capability lacking in the traditional approaches. *See* CIRCUIT (ELECTRONICS); DIGITAL-TO-ANALOG CONVERTER; LOGIC CIRCUITS; MICROPROCESSOR.

Philip V. Lopresti

Waveform

The pictorial representation of the form or shape of a wave, obtained by plotting the amplitude of the wave with respect to time. There are an infinite number of possible waveforms (see **illus.**). One such waveform is the square wave, in which a quantity such as voltage alternately assumes two discrete values during repeating periods of time. Other waveforms of particular interest in electronics are the sine wave and rectified sine wave, the sawtooth wave and triangular wave, and the arbitrary wave—a recurrent waveform which takes on an arbitrary shape over one complete cycle; this shape is then repeated in successive cycles.

Each of these waveforms has a shape which repeats periodically in time. It is possible to characterize any of them mathematically by a Fourier series, a weighted sum of terms consisting of the basic periodic trigonometric functions: sines and cosines. A periodic waveform thus can be represented as a constant or dc term, plus a sum of harmonically related sine and cosine terms where the sine and cosine frequencies are integral multiples of the fundamental frequency. The Fourier series is given beside each waveform in the illustration as a function of time t, where E_m is the maximum value of the wave and T is the period. *See* FOURIER SERIES AND TRANSFORMS; NONSINUSOIDAL WAVEFORM; WAVE-SHAPING CIRCUITS.



Common electrical waveforms.

Waveform determination

A waveform describes the variation of a quantity with respect to time. The necessary measurements to determine a waveform are normally carried out and

presented in one of two ways: the amplitude may be presented as a function of time (time domain), or an analysis may be given of the relative amplitudes and phases of the frequency components (frequency domain). Although the simplest instruments measure and display the information in the same domain, it is possible to convert the results from the time domain to the frequency domain and vice verse by mathematical processing.

Waveforms may be divided into two classes, depending on whether the signal is repeated at regular intervals or represents a unique event. The former signal is defined as a periodic or continuous wave, the latter as an aperiodic signal or transient. *See* ELECTRIC TRANSIENT; NONSINUSOIDAL WAVEFORM; WAVEFORM.

Time domain. The oscilloscope is an example of an instrument that measures and displays directly in the time domain, by deflecting an electron beam in a vertical direction in accordance with the signal while scanning at a uniform rate in the horizontal direction. The position of the beam is revealed by a fluorescent screen. *See* GRAPHIC RECORDING INSTRUMENTS; OSCILLOSCOPE.

Frequency domain. Several methods may be used to obtain the spectral content of a waveform. In the simplest, the signal is applied to a filter that is manually tuned in turn to each frequency that is expected to be present. In order to automate the measurement, the tuning of a filter may be varied by a linear, logarithmic, or other sweep and the resulting output displayed. However, any transient signal that occurs at a frequency that is different from the filter frequency will be missed. This is an important restriction that limits the technique to continuous waveforms. Specific instruments that display the frequency components of the signal are often called spectrum analyzers or harmonic analyzers. *See* ELECTRIC FILTER; HETERODYNE PRINCIPLE; SPECTRUM ANALYZER.

For applications such as the analysis of music and speech, it is necessary to simultaneously measure the intensity variation at a number of frequencies. Instead of the swept filter, an array of separate fixed tuned filters may be used, each adjusted to respond to a slightly different frequency. The amplitude of the signal in each filter is sampled in turn and displayed, giving a histogram of the frequency components of the waveform. This approach, though simple in concept, has largely been replaced by digital techniques, which are simpler to implement and offer greater flexibility. The time variation of the signals at each frequency can be analyzed as a separate waveform; this is called spectral analysis.

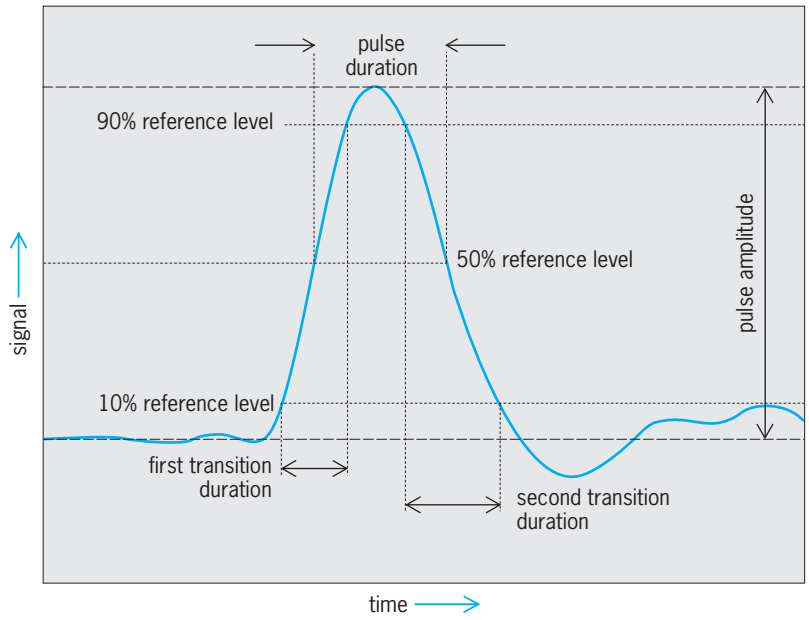
Spectral analysis extends analysis of signals in the frequency domain to nonperiodic and other signals by introducing the concept of power spectral density (PSD), in which the power estimates are displayed as a continuous function of frequency instead of being limited to discrete harmonics. The data collected during a given time interval either can be used to provide the mean power spectral density over the whole period or can be subdivided to show how the power spectral density is varying with time. In the latter case, the power spectral density estimates will be less accurate. A further extension of the technique allows additional dimensions to the array of data so that the technique can be applied to, for example, image analysis. These modern techniques of waveform determination provide the crucial key to advances in fields as widely separated as the identification of radar echoes, biomedical signal analysis, speech processing, and the transmission of audio and video signals over links of very low bandwidth. *See* DATA COMPRESSION; IMAGE PROCESSING; SIGNAL PROCESSING; SPEECH RECOGNITION.

Digital techniques. Many modern instruments use techniques in which the signal to be measured is sampled and digitized. The key elements of digital oscilloscopes are a high-speed analog-to-digital converter, sufficient high-speed memory to store the results, and a display. Because the signal is continuously digitized and stored in memory, these instruments are ideal for measuring transient waveforms. As the data can be read out of digital storage at any convenient speed, low-bandwidth display circuits are adequate. *See* ANALOG-TO-DIGITAL CONVERTER.

Once the data have been collected in digital form, they can be processed in many different ways. The application of a discrete Fourier transformation (DFT) to the amplitude data enables the information to be presented in the frequency domain. In this way, harmonic distortion that was invisible on a directly displayed waveform can be made obvious. The fast Fourier transform (FFT) is a method by which the arithmetic involved is simplified and thus accelerated. Wavelet transforms are another powerful technique that can be used to analyze waveforms. These are well suited to showing localized intermittent periodicities such as natural events and transients. *See* FOURIER SERIES AND TRANSFORMS; SPEECH RECOGNITION; WAVELETS.

Typical desktop computers can carry out fast Fourier transform transformations on 1024 data points in less than 0.1 s. This allows transformations to be carried out on speech signals in real time. Some applications may need more processing power than is available with a typical desktop computer. Digital signal processor and field-programmable gate array integrated circuits can carry out fast Fourier transforms at much higher speeds. The best can carry out transformations on 1024 data points at sample rates of over 350 kHz (in 2.85 μs). The availability of such computing power makes it practical to design instruments that collect the data in the most convenient way and display the results in the form most useful to the user. *See* LOGIC CIRCUITS.

Parametric representation. Often, waveforms contain too much detailed information for easy comparison. Instead, they can be represented by simple parameters, such as the pulse duration or transition duration. Because it is important for business that everyone agrees common definitions, the Institute of Electrical and Electronic Engineers has written a standard which sets out definitions of all these parameters and how they should be measured. The transition durations (also commonly known as rise



Typical pulse waveform showing the definitions of the transition duration and pulse duration parameters.

time and fall time) are measured between the 10% and the 90% pulse amplitude reference levels. The pulse duration (also commonly known as full width at half maximum or FWHM) is measured at the 50% reference level. The pulse duration and transition durations are the key parameters often used to characterize a pulse waveform (see **illustration**).

Very short pulses. All instruments used for waveform determination have a finite temporal resolution, and this is related to the instrument bandwidth. A larger-bandwidth instrument will have a shorter temporal resolution and vice versa. The temporal resolution is defined in terms of the pulse duration of the recorded waveform for an input pulse of infinitesimally short duration. How much the waveform is distorted by this temporal response depends on the difference between the response time of the instrument and the duration of the artifacts within a pulse. Equivalently, it depends on whether the spectral content of the pulse includes frequencies that are higher than the maximum to which the instrument is sensitive, its bandwidth. Very short pulses are loosely defined as those for which the available measurement instruments will significantly distort the waveform. Thus, whether or not a pulse is said to be very short depends on the available measurement instruments. Pulses with durations within one order of magnitude of the temporal response of the instrument might be considered to be distorted significantly. In such cases, the duration is specified by the time between the points of half the maximum signal, the pulse duration (see illustration).

The oscilloscope is widely used for recording electrical pulses. Sampling oscilloscopes can be used only for repetitive pulses, and transient pulses are recorded by real-time oscilloscopes and transient digitizers, which are about five times slower. In a sampling oscilloscope, a sufficiently large number of samples is taken to define the waveform. The sampling rate is not necessarily high and a different part of the waveform is measured for each trigger event.

Optical pulses can be converted to electrical pulses by devices such as a photodiode and recorded by the oscilloscope. Where a photodiode is used, the combined response of the photodiode and recording instrument must be considered. Instruments that combine a photodiode and a sampling oscilloscope together are used to measure optical communications data signals. *See* OPTICAL DETECTORS.

Very short pulses for which the waveform can be unambiguously determined by a sampling oscilloscope include electrical pulses where the durations of artifacts within the pulse are at least about 5 picoseconds. Electric fields in microwave circuits with frequencies from gigahertz to terahertz can be measured using electro-optic sampling. The electric field of a circuit changes the birefringence of an electrooptically active crystal, modifying its behavior in transmitting optical radiation. The change in birefringence is probed using a femtosecond laser, resulting in time resolutions in the subpicosecond range. *See* BIREFRINGENCE; ELECTROOPTICS; LASER.

Typically, optical pulse widths of less than 10 femtoseconds can be generated in the laboratory by commercial laser systems. To give an idea of scale, an 85-fs light pulse is about 25 μm (0.001 in.) long. These short light pulses can be used for many applications such as absolute frequency standards, to generate terahertz radiation, or to generate even shorter pulses in the extreme ultraviolet region. The durations of these pulses are estimated by autocorrelation. This technique does not give the pulse shape but only the FWHM for an estimated pulse shape. An extension of the autocorrelation technique allows the shapes of optical pulses of subpicosecond duration to be measured using the frequency-resolved optical gating (FROG) technique. The method is analogous to an oscilloscope but almost a thousand times faster. *See* FREQUENCY MEASUREMENT; OPTICAL PULSES; SUBMILLIMETER-WAVE TECHNOLOGY; ULTRAFAST MOLECULAR PROCESSES.

The streak camera is the fastest instrument used to record pulse shape directly. It is an electron imaging tube that converts an optical pulse to a distribution of intensities across a phosphor screen, recording the pattern using a photodiode array. The response of the streak camera is presently limited to about 500 ps.

Often the technology used to generate a short pulse is common to that of the measurement instrument, so that both are limited to a similar response. The resulting waveform is the convolution of the pulse shape with the instrument response. Mathematical techniques can be used to reduce this distortion, depending on the extent to which the response of the instrument is known or calculable. This process is called deconvolution.

David A Humphreys

Bibliography. R. N. Bracewell, *The Fourier Transform and Its Applications*, 3d ed., McGraw-Hill, 1999; *IEEE Standard on Transitions, Pulses and*

Related Waveforms, IEEE Std 181, 2003; S. M. Kay, *Modern Spectral Estimation: Theory and Application*, Prentice Hall, 1988, paper 1999; L. S. Marple, *Digital Spectral Analysis with Applications in C, FORTRAN, and MATLAB*, Prentice Hall, 1987; N. S. Nahman, Picosecond-domain waveform measurements, *Proc. IEEE*, 66:441–454, 1978; S. Prentiss, *The Complete Book of Oscilloscopes*, 2d ed., 1992.

Waveform generator

A term usually applied to an instrument that generates voltage waves using digital technology. These digital instruments have largely supplanted the analog instruments that traditionally served in this role. Digital instruments have a number of important advantages over analog ones: they are more flexible and can be used to generate waveforms that are beyond the capabilities of analog instruments, they produce more accurate replicas of the desired mathematical waveforms, and their output is more stable and predictable over time. *See* FUNCTION GENERATOR; SIGNAL GENERATOR; WAVE-SHAPING CIRCUITS.

The basic components of these instruments are: a digital signal memory that contains numeric samples of the desired signal waveform, a sequencing system that accesses this memory in a predetermined pattern, a high-accuracy digital-to-analog converter that transforms the contents of the accessed memory location to proportional analog voltages, a low-pass filter to smooth the output of the converter, and a digital clock—the timing device that sets the memory access rate. *See* DIGITAL-TO-ANALOG CONVERTER; ELECTRIC FILTER; SEMICONDUCTOR MEMORIES.

The two most common approaches to implementing these instruments are direct digital synthesis (DDS) and arbitrary waveform generation (AWG). Each has its own advantages and each plays an important role. Direct digital synthesis typically is used to generate high-accuracy versions of standard analog signal waveforms: sine, triangular, and square waves. Instruments that employ arbitrary waveform generation are much more flexible and can generate a repetitive signal waveform that takes on values limited only by the word size of each memory cell and by the number of memory locations in the instrument. Instruments that use direct digital synthesis can be programmed to change their output from one form to another essentially instantaneously, while devices that use arbitrary waveform generation are limited by the time it takes to replace their signal memory contents with new signal patterns.

Direct digital synthesis. Direct digital synthesis employs a fixed-rate digital-to-analog converter followed by a high-quality fixed-frequency low-pass filter. A phase increment register and a phase accumulator control the output frequency of the generator. To illustrate the concept, note that one full cycle of a sine wave is traversed as its angle covers 360° and assume that consecutive memory locations hold values of the sine function in one-degree increments: sin(0°), sin(1°), . . . , sin(359°). Sequentially scanning and converting these memory locations will take 360 "ticks" of the timing clock to generate one cycle of the output wave and the resulting frequency would be 1/360 the frequency of the clock. The phase increment register sets the memory step size and the phase accumulator increases by this amount for each clock tick. Thus, the accumulator contains a memory address representing the angle and the addressed memory cell contains the value of the sine of the angle. The direct digital synthesis memory bank is often called a look-up table since the approach is to "look up" values of the function to be generated. Since the sine of 360° is identical to the sine of 0°, and since the function repeats periodically every 360°, the accumulator is programmed to automatically return to 0° when its angle tries to pass 359°, and the conversion process repeats indefinitely.

For a 360-MHz clock and an increment size of 1°, the signal frequency will be 1 MHz. Increasing the

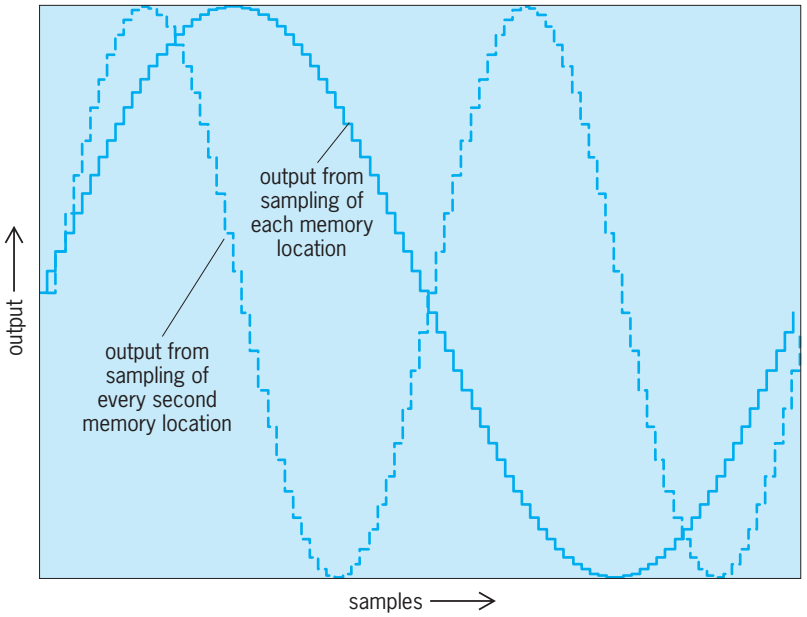


Fig. 1. Stepped-response output of direct digital synthesis (DDS) digital-to-analog converter.

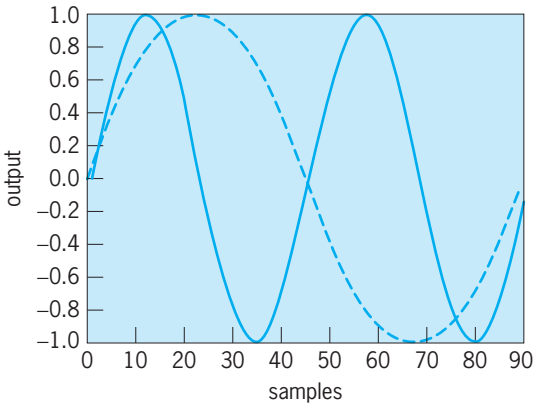


Fig. 2. Direct digital synthesis (DDS) output smoothed after low-pass filtering.

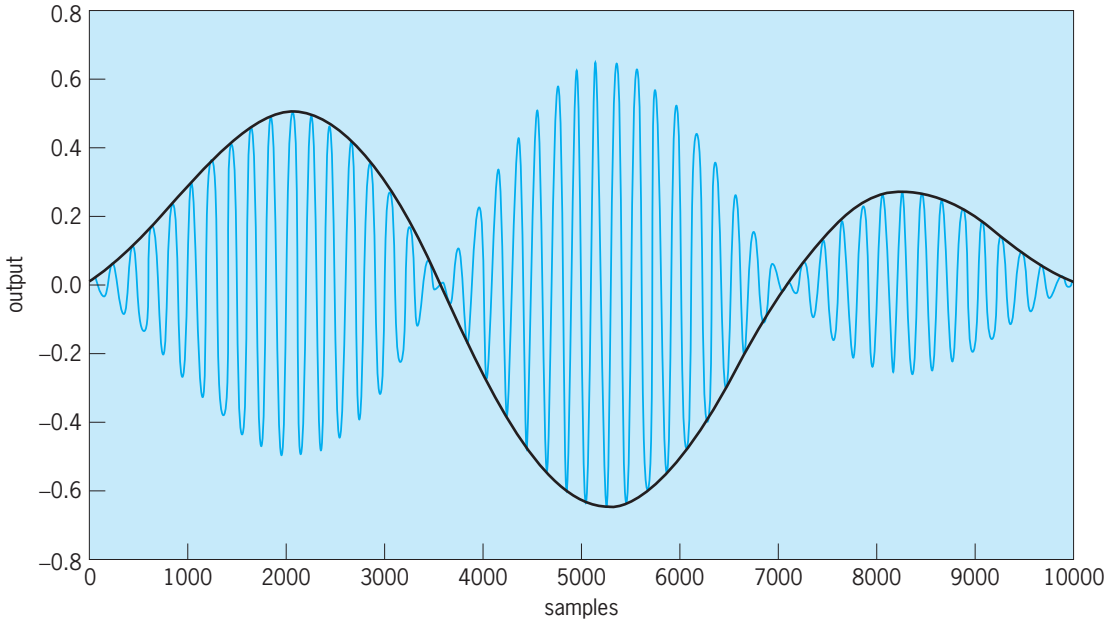


Fig. 3. Sinusoidal carrier modulated by complex pulses.

phase increment register value to 2° will access every second memory location, the sine wave will complete one cycle in 180 clock "ticks," and the signal frequency will be 1/180 the frequency of the clock, or 2 MHz in this case.

In practice, the minimum phase increment available for an instrument is much smaller than 1°. Some high-performance generators contain phase increments smaller than 1 billionth of a cycle.

The concept is illustrated in **Figs. 1** and **2**. Figure 1 shows the output of the digital-to-analog converter for a 90-sample sine wave look-up table, where each successive location contains the sine of an angle 4° greater than the preceding one. Sampling each memory location in sequence generates the first wave and sampling every second location generates the second. The steps in the figure correspond to the increments in the digital-to-analog converter output as each memory location is accessed. Figure 2 shows the smoothed, low-pass filtered output for each generated frequency. The physical output frequency depends on the timing clock frequency.

Arbitrary waveform generation. Arbitrary waveform generation is conceptually simpler than direct digital synthesis. As in direct digital synthesis, numerical values for a waveform are stored in consecutive memory locations, but the locations are always accessed sequentially with no skipping. The values inserted in each location can be arbitrary—hence the name arbitrary waveform generator. As in the direct digital synthesis case, the waveforms are repetitive, returning to the start point as the end of the signal memory is reached.

Sine waves and the rest of the waveforms that can be generated by direct digital synthesis methods also can be generated by an arbitrary waveform generator, but this is not the principal use for these instruments. Their ability to generate almost any conceivable periodic waveform is usually used to generate waveforms with much more complex shapes. These values can be chosen from preconfigured signals that come with the instrument, extracted from real-world data via analog-to-digital conversion, copied from other programs such as spreadsheets, or entered point-by-point by hand.

They can, for example, mimic the behavior of the quadrature amplitude modulation (QAM) systems that form the backbone of most digital communications links. A two-channel arbitrary waveform generator can generate QAM signals consisting of a sine waveform with amplitude and phase variations on one channel and a cosine waveform with separate amplitude and phase variations on the other, combine these waves, and distort them in ways that correspond to imperfect transmitter hardware or to poor communications links. The arbitrary waveform generator thus can generate the input signals needed to submit a link receiver to carefully controlled stress tests. *See* AMPLITUDE MODULATOR; DATA COMMUNICATIONS; MODULATION.

Figure 3 illustrates an example of how an arbitrary waveform generator might generate a complex signal, in this case the amplitude modulation of the sinusoidal carrier of a QAM transmitter by a sequence of pulses. Such a waveform cannot be generated by an analog function generator or by a waveform generator based on direct digital synthesis.

Philip V. Lopresti

Bibliography. B.-G. Goldberg, *Digital Frequency Synthesis Demystified*, LLH Technology, Eagle Rock, VA, 1999; J. H. Reed, *Software Radio: A Modern Approach to Radio Engineering*, Prentice-Hall, Upper Saddle River, NJ, 2002.

Waveguide

A device that constrains or guides the propagation of electromagnetic radiation along a path defined by the physical construction of the guide. Electromagnetic

waves may propagate in space, as radio waves, but for many purposes waves need to be guided with minimum loss from the generating point to a point of application. Several guiding systems are of importance, including two-conductor transmission lines, various forms of striplines used in microwave integrated circuits, hollow-pipe waveguides, and dielectric waveguides. Since transmission lines and striplines are discussed in other articles, the emphasis in this article is on hollow-pipe and dielectric guides. Hollow-pipe guides are used primarily in the microwave region of the spectrum, dielectric guides primarily in the optical region. *See* COAXIAL CABLE; ELECTROMAGNETIC WAVE TRANSMISSION; MICROWAVE; TRANSMISSION LINES.

Hollow-Pipe Waveguides

Hollow-pipe waveguides consist of a dielectric region, usually air, surrounded by a closed good conductor such as silver, copper, aluminum, or brass. The cross-sectional shape is usually rectangular (**Fig. 1**), but may be circular or of a variety of other shapes. Voltage and current concepts, so useful for transmission lines, are not so useful for waveguides. Distributions of electric and magnetic fields, obtained from Maxwell's equations, are needed. *See* MAXWELL'S EQUATIONS.

Solution of the field equations shows that there is an infinite number of modes for these guides, where a mode is a solution that maintains its transverse pattern but attenuates and shifts in phase as it propagates along the guide. These modal patterns may be likened to the resonant modes of a drumhead, with low-order modes having only a few transverse variations and high-order modes having many.

All modes in these guides have cutoff properties. Below a certain critical frequency, the mode will not propagate but only attenuates if excited; above the cutoff frequency, though, it may propagate with a finite phase velocity and small attenuation (zero attenuation for perfect dielectric and conductor). Higher-order modes usually have higher cutoff frequencies.

If the guiding pipe is perfectly conducting, the modes may be divided into transverse magnetic (TM) and transverse electric (TE) types. For the former, the magnetic field is confined to the transverse plane, while the electric field is so confined for the latter. These classifications remain useful for the practical metallic conductors used for such guides.

Principal mode in rectangular. The principal mode in a guide of rectangular cross section (Fig. 1), with air dielectric, constitutes the most commonly used guide and mode. The electric field of the principal mode has only a vertical component with no variations vertically, and a half-wave sinusoidal variation in the transverse (x) direction. The magnetic field has only transverse (x) and axial (z) components, forming closed loops surrounding the time-varying electric field (Fig. 1). The current flow on the interior of the metal walls is normal to the tangential magnetic field at the walls and has both axial and transverse components. The electric field is confined to the transverse plane, so the wave is a transverse electric type. Since the electric field has

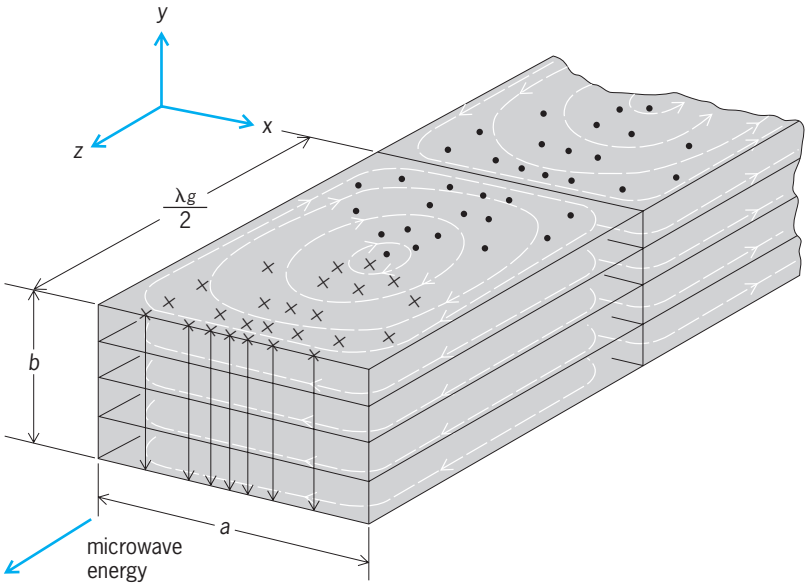


Fig. 1. Rectangular hollow-pipe waveguide with field patterns for a propagating TE_{10} mode. Solid lines indicate electric field and broken lines magnetic field; ×'s indicate energy *E* pointing down; solid circles indicate *E* pointing up. (*After MIT Radar School Staff, Principles of Radar, 1952*)

one variation in x and none in y, the designation is TE_{10}.

For a given guide of width a, the mode will propagate average power only for frequencies higher than a cutoff frequency f_c, which corresponds to a frequency for which the width is just a half-wavelength. For frequencies below f_c, the longer half-wavelength will not "fit" in the guide and, if excited, will only attenuate exponentially with z, without any average power transfer. Thus, hollow-pipe waveguides used for power transfer must have transverse dimensions that are of the order of wavelength, so they are most used in the microwave region of the spectrum.

The expressions for the fields of this TE_{10} mode, obtained from Maxwell's equations subject to the boundary conditions of the walls (taken here as perfectly conducting), are given by Eqs. (1), (2), and (3).

$$E_y = E_0 \sin\frac{\pi x}{a} \cos\omega\left(t - \frac{z}{v_p}\right) \tag{1}$$

$$H_x = -\frac{E_0}{\mu v_p} \sin\frac{\pi x}{a} \cos\omega\left(t - \frac{z}{v_p}\right) \tag{2}$$

$$H_z = -\frac{\pi E_0}{\omega\mu a} \cos\frac{\pi x}{a} \sin\omega\left(t - \frac{z}{v_p}\right) \tag{3}$$

Here a is the guide width; $\omega = 2\pi f$, where f is the frequency; μ is the permeability, equal to $4\pi \times 10^{-7}$ henry per meter for nonmagnetic dielectrics; E_0 is the electric field strength in volts per meter; and v_p is the phase velocity in the axial direction. This velocity is given by Eq. (4), where c is the velocity of light

$$v_p = \frac{c}{\sqrt{\epsilon_r}} \left[1 - \left(\frac{f_c}{f}\right)^2\right]^{-1/2} \tag{4}$$

in vacuum; ϵ_r is the relative permittivity or dielectric

constant; and f_c is the cutoff frequency, equal to $c/(2a\sqrt{\epsilon_r})$. Equation (4) shows that the phase velocity becomes imaginary, indicating no real propagation of energy, for $f < f_c$. The guide wavelength λ_g for $f > f_c$ is defined as the axial distance for which a guided mode changes phase by 2π, and is a useful concept. For the TE_{10} mode, and a nonmagnetic dielectric, it is given by Eq. (5), where λ_0 is the free-space wavelength.

$$\lambda_g = \frac{v_p}{f} = \frac{\lambda_0}{\sqrt{\epsilon_r}}\left[1 - \left(\frac{f_c}{f}\right)^2\right]^{-1/2} \tag{5}$$

See PHASE VELOCITY.

The axial power flow of the mode is obtained from the transverse field components E_y and H_x from Poynting's theorem and is found to be given by Eq. (6). This also becomes imaginary when v_p is imaginary, that is, for $f < f_c$ *See* POYNTING'S VECTOR.

$$P = \frac{abE_0^2}{4\mu v_p} \qquad \text{watts} \tag{6}$$

The maximum power transfer capability of this guide is limited by the breakdown electric field at the guide center ($x = a/2$). Taking air breakdown as 15 kV/cm, the power limit for a guide at 10 GHz would be about 260 kW.

Power or energy is said to travel at a velocity v_g, called group velocity, given for a nonmagnetic dielectric by Eq. (7). Group velocity is actually the velocity

$$v_g = \frac{c}{\sqrt{\epsilon_r}}\left[1 - \left(\frac{f_c}{f}\right)^2\right]^{1/2} \tag{7}$$

of the envelope of a modulated wave, so it could represent the envelope of a short pulse of the microwave signal. Both v_p and v_g are functions of frequency, so these guides are dispersive and the pulse spreads or changes shape as it propagates. *See* GROUP VELOCITY.

Higher-order modes in rectangular guides. The waveguide of rectangular cross section has other modal solutions (**Fig. 2**). TM_{21}, for example, denotes a wave with magnetic field confined to the transverse plane, having two half-sine variations with x and one with y. The cutoff frequency for either the TM or TE mode with m variations in x and n with y, for a nonmagnetic dielectric, is given by Eq. (8). (The numbers m or n, but not both, can

$$(f_c)_{mn} = \frac{c}{\sqrt{\epsilon_r}}\left[\left(\frac{m}{2a}\right)^2 + \left(\frac{n}{2b}\right)^2\right]^{1/2} \tag{8}$$

be zero for transverse electric waves, but both must be nonzero for transverse magnetic waves.) The cutoff frequency increases with m and n (for a given guide size), and, if $a > b$, that for the TE_{10} mode is the lowest.

It is usually desirable to select the guide size so that only one mode propagates, so that there is not interference among several modes propagating with different phase velocities. For example, for the microwave X-band ($f \sim 10$ GHz), a standard guide size (interior dimension) is 0.9 by 0.4 in. (2.3 by 1.0 cm). The cutoff frequency for the TE_{10} mode is then 6.56 GHz so that $f > f_c$ for this mode; for the TE_{01} mode it is 14.76 GHz, and it is 16.15 GHz for TE_{11} or TM_{11}. The higher-order modes may be excited at the beginning of the guide, and at discontinuities, but if $f < f_c$ they are localized and do not propagate. Their reactive energy may produce reflections or other circuit effects that must be taken into account in the matching process.

Although the expressions given are for guides with ideal conductors and dielectrics, good conductors and dielectrics do not change these much but do produce an attenuation of the mode as it propagates. Most guides are air-filled, and typical attenuations for the TE_{10} mode in a rectangular guide are 0.075 dB/ft (0.25 dB/m) at 10 GHz.

Circular hollow-pipe waveguides. For some applications, the symmetry of a circular-cylindrical guide is desirable. These also have an infinite number of mode types that may be divided into transverse electric and transverse magnetic classes when losses are small. Cutoff frequencies for the modes of circular

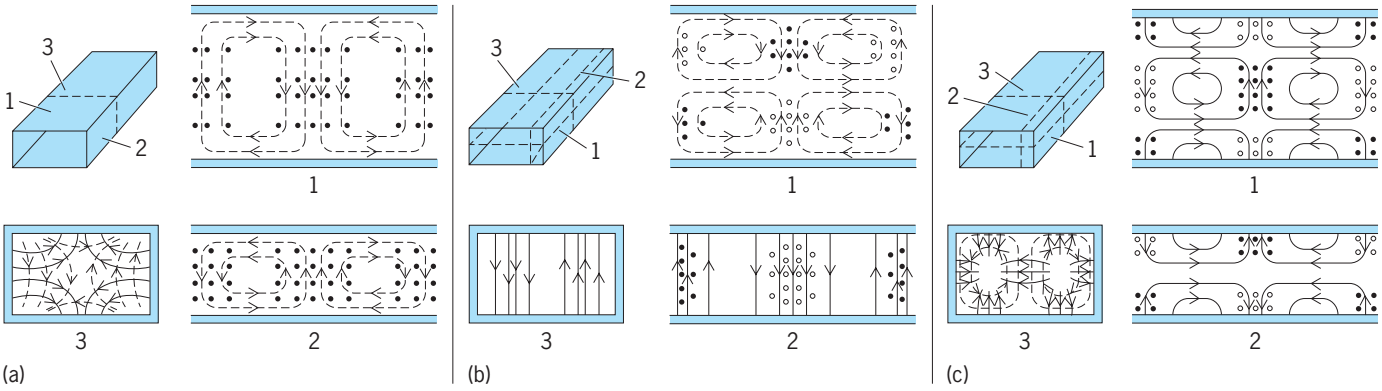


Fig. 2. Various modes of rectangular hollow-pipe guide. Picture of waveguide in each part of figure shows orientations of waveguide walls or cross sections, identified as 1, 2, and 3, on which fields are shown. Solid lines indicate electric field and broken lines magnetic field; circles indicate field perpendicular to plane of diagram, with solid circles indicating field pointing out of page and open circles indicating field pointing into page; circles accompanying solid lines indicate magnetic field, and those accompanying broken lines indicate electric field. (*a*) TE_{11} mode. (*b*) TE_{20} mode. (*c*) TM_{21} mode. (*After S. Ramo, J. R. Whinnery, and T. Van Duzer, Fields and Waves in Communication Electronics, 3d ed., John Wiley and Sons, 1994*)

guides with a nonmagnetic dielectric are given by Eqs. (9) and (10), where a is radius of the cylinder, p_{nl}

$$(f_c)_{\mathrm{TM}_{nl}} = \frac{p_{nl}c}{2\pi a\sqrt{\epsilon_r}} \tag{9}$$

$$(f_c)_{\mathrm{TE}_{nl}} = \frac{p'_{nl}c}{2\pi a\sqrt{\epsilon_r}} \tag{10}$$

represents the lth zero of the nth-order Bessel function, and p'_{nl} is the lth zero of its derivative. The first subscript in the designations TE_{nl} or TM_{nl} denotes the number of variations circumferentially, and the second subscript denotes the number of variations radially. The principal mode, or mode with longest cutoff wavelength (lowest cutoff frequency), is the TE_{11} mode, which, apart from the curvature required by the boundaries, is analogous to the principal TE_{10} mode of rectangular guides. The TE_{11} circular mode, however, has two polarizations with the same cutoff wavelength (called degenerate modes). The two may be excited in phase quadrature to produce fields with circular polarization, which may be desirable in certain rotating joints.

A second mode of special interest is the TE_{01} mode (or any TE_{0l} mode), in which the attenuation from conducting losses in the walls decreases with increasing frequency. It was consequently considered promising as a medium for long-distance, broadband communication, but the simpler optical fiber has proved the appropriate medium for these broadband links.

Other cross sections. An infinite number of other guide shapes is possible. A guide of elliptic cross section eliminates the degeneracy of modes in the circular guide, giving mode patterns somewhat between those of the circular and rectangular guides. A rectangular guide with a central ridge (called a ridge waveguide) gives a lower cutoff frequency than the rectangular guide with the same width, but also has lower power-handling capabilities because of the narrower gap in the region of highest field.

Circuit elements. A variety of circuit elements is required for exciting desired modes, filtering, coupling to passive and active elements, and other necessary networking functions. Excitation of a particular mode may be by probes, along the direction of the mode's electric field, by loops normal to magnetic field lines, or by the charge streams of an active vacuum-tube or semiconductor device placed within the guide.

A thin diaphragm introduced from the top of the guide (or the bottom, or both) may be represented as a capacitive shunt element in a transmission-line model. A diaphragm introduced from the side acts as an inductive shunt element. Circuit representations for a variety of probes, posts, irises, and other waveguide elements have also been analyzed and tabulated. Combinations of these elements may then be used to synthesize filters, reflectors, or matching elements in the guide.

Other important waveguide elements are the directional coupler, isolator, and magic-T. In the directional coupler, there is coupling to an auxiliary guide

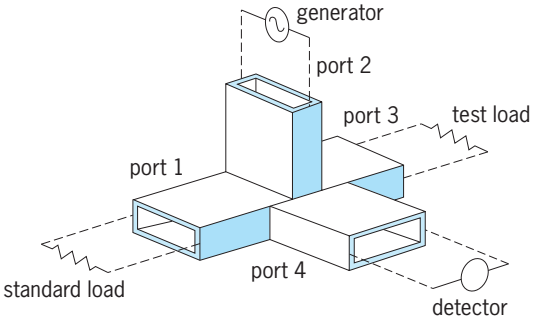


Fig. 3. Magic-T network in rectangular hollow-pipe waveguide used as a bridge. (*After S. Ramo, J. R. Whinnery, and T. Van Duzer, Fields and Waves in Communication Electronics, 3d ed., John Wiley and Sons, 1994*)

in such a way that the output of one of its ports is proportional to the wave traveling in the forward direction, and the output of the other is proportional to the reverse wave. The isolator makes use of the nonreciprocal properties of ferrites with an applied magnetic field to pass the forward-traveling wave of the guide but to eliminate the reflected wave. The magic-T (**Fig. 3**), and other forms of the microwave hybrid network, gives no output in port 4 when ports 1 and 3 are balanced, so it may be used in a bridge arrangement to compare a test load with a standard load. It is especially useful in the balanced mixers of heterodyne systems. E-plane T junctions, H-plane T junctions, and Y junctions are simple forms that are used in branching networks. *See* DIRECTIONAL COUPLER; GYRATOR.

Dielectric Waveguides

A dielectric waveguide consists of one dielectric material, called the core, surrounded by a different dielectric, called the cladding. The permittivity (dielectric constant), or refractive index, of the core is larger than that of the cladding, and under proper conditions electromagnetic energy is confined largely to the core through the phenomenon of total reflection at the boundary between the two dielectrics. *See* PERMITTIVITY; REFLECTION OF ELECTROMAGNETIC RADIATION; REFRACTION OF WAVES.

Early dielectric guides were so lossy that they could be used only over short distances. Dielectric light pipes found surgical and laboratory use, and dielectric rods, called polyrod antennas, were used as radiators in certain World War II microwave radars. In 1969 silica fibers were developed with attenuations of 32 dB/mi (20 dB/km), low enough to be of use for optical communication applications. Since then, further improvements have reduced losses to as low as 0.3 dB/mi (0.2 dB/km). Fiber guides are now the basis for a worldwide optical communication network. *See* FIBER-OPTICS IMAGING.

Planar, rectangular, and thin-film forms of dielectric guides are also important in guiding optical energy from one device to another in optoelectronic and integrated optic devices, for example, from a semiconductor laser to an electrooptic modulator on a gallium arsenide substrate. Because of the simpler geometry, the planar forms will be used to explain

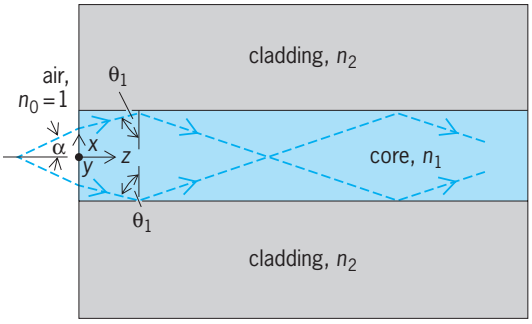


Fig. 4. Planar dielectric waveguide with ray paths that make up guided waves when angle θ_1 is greater than the critical angle for total reflection. Cladding refractive index n_2 is smaller than core refractive index n_1. Transverse direction (*y* direction) is perpendicular to the plane of the diagram.

the principle. *See* INTEGRATED OPTICS; LASER; OPTICAL MODULATORS.

Planar dielectric guides. The principle of dielectric guiding of electromagnetic energy can be illustrated in the symmetrical dielectric slab guide (**Fig. 4**), in which a slab of core material with refractive index n_1 is placed between two slabs of cladding with a refractive index n_2 that is smaller than n_1. Rays (normal to the wavefronts of plane waves) will be refracted in going from air to the core, and then reflected at the interface between core and cladding. If the angle of incidence θ_1 is greater than the critical angle for total reflection from this surface, all energy is reflected. The critical angle θ_c is given by Eq. (11). For angles

$$\theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right) \tag{11}$$

θ_1 greater than θ_c, the energy is thus trapped by the guiding system with no average power leaking out; for steeper angles ($\theta_1 < \theta_c$), there is some transmission to the outer medium on each reflection, representing a transmission loss to the propagating wave. Thus, the condition $\theta_1 > \theta_c$ is required for guiding without leakage or radiation loss.

In tracing the rays from the air at the input to the guide, there is some maximum angle α for rays to satisfy the above condition for guiding. The sine of this angle is the numerical aperture (NA) and is given by Eq. (12). Only rays within the wedge of angle $2\,\alpha_{\max}$

$$\text{NA} = \sin\alpha_{\max} = \sqrt{n_1^2 - n_2^2} \tag{12}$$

satisfy the condition for total reflection. Because of the square-root relationship, even small differences between n_1 and n_2 yield reasonable values of the numerical aperture. For example, if $n_1 = 1.50$ and $n_2 = 1.49$, NA $= 0.173$ and $\alpha_{\max} \approx 10^0$.

The ray picture is clear and accurate, but the modes of this guiding system can also be found by obtaining appropriate solutions of Maxwell's equations. If variations in the transverse direction are negligible, these divide into transverse electric and transverse magnetic types, as in the hollow-pipe guides. The field does extend into the cladding dielectric (**Fig. 5**), but is evanescent (that is, it dies off exponentially with distance away from the interface). A parameter v is useful in telling which modes are guided and which radiate into the cladding. This parameter is defined by Eq. (13), where $2a$ is the height of the

$$v = \frac{2\pi a}{\lambda}\,\text{NA} \tag{13}$$

slab, λ is the free-space wavelength, and the numerical aperture NA is defined by Eq. (12). The TE_0 mode is guided for all values of v; but higher-order modes, TE_m or TM_m, are guided only above some cutoff v defined by Eq. (14). To increase the v parameter of a

$$(v)_{\text{cutoff}} = \frac{m\pi}{2} \tag{14}$$

mode so that it is above its cutoff value, it is possible to increase the slab height a, increase the frequency (decrease λ), or increase NA by making the difference between n_1 and n_2 greater.

Modes in the zigzag ray picture are defined as those rays for which the phase shift along the ray, including the phases of the reflections, is a multiple of 2π after a complete path of double reflection.

Rectangular and quasirectangular guides. For applications to integrated optics, in which the dielectric guide is used to direct optical energy from one component to another, it is necessary to have some lateral confinement as well as the vertical confinement provided by the layers of the slab guide (Fig. 4). The lateral confinement is provided by surrounding the guiding core with a dielectric of lower refractive index on the sides as well as on top and bottom, or processes such as ion implantation or diffusion may produce an inhomogeneous region of higher index than the surroundings to effect the confining region. *See* ION IMPLANTATION.

Optical fibers. By far the most important dielectric guide at present is the optical fiber used for optical communications. Here the round core is surrounded by a cladding of slightly lower refractive index. The combination is surrounded by a protective jacket to prevent corrosion and give added strength, but this jacket plays no role in the optical guiding. The principle of guiding is the same as that described for the slab guide. From the ray point of view, rays are totally reflected from the boundary between n_1 and n_2

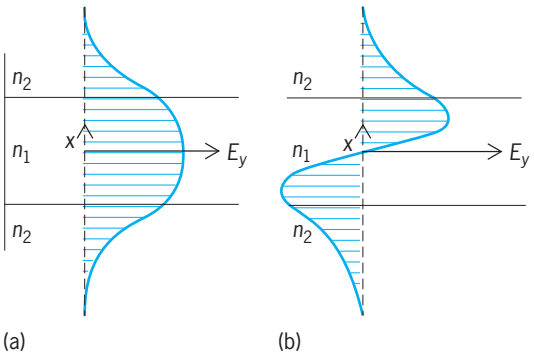


Fig. 5. Distributions of electric field E_y versus coordinate distance normal to planar dielectric guide axis for (*a*) lowest-order ($m = 0$) and (*b*) first-order ($m = 1$) transverse electric modes.

for angles flatter than the critical angle defined by Eq. (11). Numerical aperture (NA) is here also a useful concept and is given by Eq. (12).

Solutions of Maxwell's equations in circular cylindrical coordinates yield the modes of this guiding system. These divide into transverse electric and transverse magnetic classes only for modes that have no circumferential variation. Other modes have all field components and are called hybrid modes, HE_{mn} or EH_{mn}, depending respectively upon whether energy is predominantly magnetic or electric. The principal mode is the HE_{11}, which has no cutoff frequency. The cutoff frequency of other modes is determined by the υ parameter of Eq. (13), with a interpreted as the radius of the core.

Losses. Fibers used in optical communications are almost universally of silica (silicon dioxide), possibly with germanium or other ions as dopants. The very low attenuations achieved came about by removal of undesired impurities. The lowest values of attenuation are about 0.3 dB/mi (0.2 dB/km), occurring for a wavelength near 1.5 micrometer. This is near the theoretical limit for Rayleigh scattering. Another local minimum of about 0.8 dB/mi (0.5 dB/km) occurs for a wavelength near 1.3 μm, which is a wavelength of lower dispersion. Attenuation increases rapidly for shorter wavelengths because of the wavelength dependence of Rayleigh scattering, and also increases at wavelengths longer than 1.6 μm because of the tail of an infrared absorption band. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; SCATTERING OF ELECTROMAGNETIC RADIATION.

Polymeric or plastic optical fibers have larger losses than silica but have found use in short-distance data-link applications. There is also work on a number of materials with losses lower than silica. For example, the fluorozirconate glasses promise very low losses in the mid-infrared region of the spectrum. *See* FIBER-OPTIC CIRCUIT; OPTICAL MATERIALS.

Radiation losses may occur even in the guided regime if there are excessive bends or corrugations in the fiber. Thus it is important to design the fiber and its cabling to avoid such perturbations.

Dispersion. All dielectric guides are dispersive in that different frequency components of the signal travel at different velocities. The variation of group velocity with frequency is the relevant quantity for calculating the distortion of the modulation envelope of an analog signal, or the pulse shape of a digital signal. This quantity is called the group-velocity dispersion. It may arise in part from the variation of refractive index with frequency (called material dispersion), and in part from the basic waveguide characteristics themselves (called waveguide dispersion). For a single-mode fiber (one with the υ parameter below 2.4, so that only the HE_{11} mode is guided), material dispersion is the larger component except near a wavelength of 1.3 μm, at which point the combination becomes zero. With special dopings, or multiple claddings, this zero dispersion point can be shifted to the region of minimum attenuation, around 1.5 μm.

For the single-mode fiber, the spread in arrival time for a narrow input pulse is given by Eq. (15), where L

$$\Delta\tau = DL\,\Delta\lambda \qquad \text{picoseconds} \qquad (15)$$

is the fiber length in kilometers, $\Delta\lambda$ the wavelength spread (from source, signal, or both) in nanometers, and D a dispersion quantity given in picoseconds per kilometer of length and nanometer of wavelength spread.

For fibers whose υ parameter is greater than 2.4, more than one mode is guided, and these modes travel with different group velocities even at a single wavelength. This causes a spread in arrival time (called intermode distortion or intermode dispersion). Since modes near cutoff travel with velocities near that of the cladding, and modes well above cutoff near that of the core, this delay for length L is given approximately by Eq. (16), where c is the velocity of light in vacuum.

$$(\Delta\tau)_{\text{multimode}} \approx \frac{(n_1 - n_2)L}{c} \qquad (16)$$

Graded-index fibers. The fibers discussed so far have a sharp discontinuity of refractive index between core and cladding and are called step-index fibers. In graded-index fibers the refractive index varies within the core, decreasing with increasing radius from a maximum at the axis (**Fig. 6**). Multimode graded-index fibers have appreciably less intermode distortion than multimode step-index fibers. Index grading can also play a role in shifting the zero dispersion wavelength of single-mode fibers.

Nonlinear effects. The refractive index of silica and other glasses depends to a slight degree on the intensity of the optical signal. Although this is a small effect, it can add up over the length of a fiber to produce both undesirable and desirable effects. Intermodulation among separate signals is the most important undesirable effect. Among the useful phenomena are pulse compression and amplification using the Raman effect. Soliton propagation, in which self-phase modulation from the nonlinear

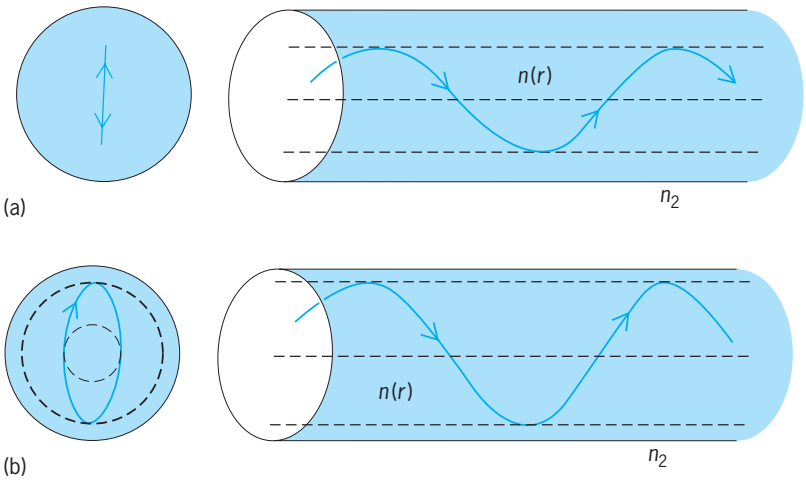


Fig. 6. Ray paths within the core of a graded-index fiber. Both end-on projection and oblique view of each path are shown. (*a*) A path through the meridian. (*b*) A skew path.

effect just compensates for group-velocity dispersion so that a short pulse propagates long distances without spreading, is especially interesting and has potential for high-data-rate communication systems. *See* NONLINEAR OPTICS; RAMAN EFFECT; SOLITON.

Fiber amplifiers. The capabilities of fiber for information transmission was greatly improved by the development of fiber amplifiers. Most important has been the silica fiber doped with erbium, providing amplification by a laser-type mechanism at a wavelength of 1.54 μm. This is the wavelength of minimum attenuation for silica fibers. Pumping of the erbium laser line is provided by semiconductor laser diodes at wavelengths of 0.98 or 1.43 μm. These amplifiers have greatly extended the transmission distance of point-to-point fiber lines and have permitted more branches in local-area uses of optical fibers. *See* OPTICAL COMMUNICATIONS; OPTICAL FIBERS.

John R. Whinnery

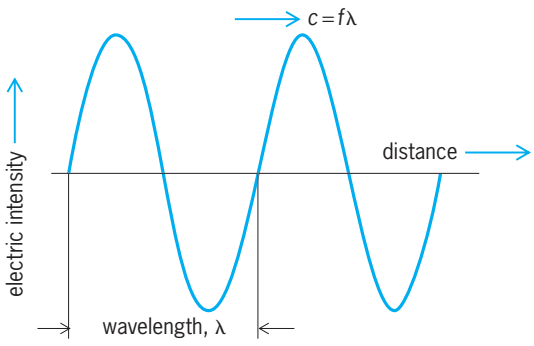
Bibliography. F. W. France, *Optical Fibre Lasers and Amplifiers*, 1991; H. A. Haus and J. R. Melcher, *Electromagnetic Fields and Energy*, 1989; D. Marcuse, *Theory of Dielectric Optical Waveguides*, 2d ed., 1991; N. Marcuvitz et al. (eds.), *Waveguide Handbook*, 1986; A. D. Olver, *Microwave and Optical Transmission*, 1992; S. Ramo, J. R. Whinnery, and T. VanDuzer, *Fields and Waves in Communication Electronics*, 3d ed., 1994; A. W. Snyder and J. D. Love, *Optical Waveguide Theory*, 1983.

Wavelength

The distance between two points on a wave which have the same value and the same rate of change of the value of a parameter, for example, electric intensity, characterizing the wave. The wavelength, usually designated by the Greek letter λ, is equal to the speed of propagation c of the wave divided by the frequency of vibration f; that is, $\lambda = c/f$ (see **illus.**). *See* WAVE (PHYSICS).

The wavelength for a sound of a given frequency varies greatly, depending upon the speed of propagation in the medium in which the sound is moving. For example, a sound wave having a frequency of 1000 Hz would have a wavelength of approximately 1 ft (0.3 m) in air, 412 ft (1.3 m) in water, and 17 ft (5.1 m) in steel. The wavelength of electromagnetic waves depends on the velocity of light in the material in which the waves are traveling. *See* WAVE MOTION.

William J. Galloway



Wavelength λ and related quantities.

Wavelength measurement

Determination of the distance between successive wavefronts of equal phase of a wave. This article discusses wavelength measurement of electromagnetic waves in the radio, microwave, infrared, and optical regions. *See* ELECTROMAGNETIC RADIATION; WAVE MOTION; WAVELENGTH.

Wavelength by frequency measurement. From the relation $\lambda = c/f$ between wavelength λ, speed c, and frequency f, the wavelength of a wave motion can be calculated if the speed is known and its frequency is measured. The ease and accuracy of electronic counting and timing make frequency measurement the most precise of all physical measurements. This method of wavelength determination is thus one of the most accurate, but only if the speed (phase velocity) is known. In free space the speed of an electromagnetic wave c_0 is, through the 1983 definition of the meter, fixed at exactly 299,792,458 m/s ($\approx$186282.397 mi/s), or roughly 300,000 km/s. Unless otherwise specified, it is general practice to quote the wavelength of an electromagnetic wave as the free-space value λ_0, given by Eq. (1).

$$\lambda_0 = \frac{c_0}{f} \tag{1}$$

See FREQUENCY MEASUREMENT; PHASE VELOCITY.

Radio and microwave regions. The presence of any dielectric material (such as air) or any magnetic matter with a permeability greater then unity will cause the wave to travel at a velocity lower than its free-space value. The speed is also altered if the waves pass through an aperture, are focused by a lens or mirror, or are constrained by a waveguide or transmission line. In such cases it may be more appropriate to measure the wavelength directly. In the pioneering experiments on radio waves, it was found that standing waves existed in space whenever reflections occurred and that these provided a convenient means of measuring the wavelength. It thus became the convention to characterize waves by their wavelength rather than by their frequency, as is now more commonly the case. Specifying the frequency is preferred because, unlike the wavelength, it is independent of the speed of propagation and does not change as the wave moves from one medium to another. *See* ELECTROMAGNETIC WAVE TRANSMISSION; TRANSMISSION LINES; WAVEGUIDE.

Interferometer methods. In the microwave region the wavelengths are sufficiently short that it is convenient to measure them by using interferometer techniques directly analogous to those used with light.

In a typical interferometer used in the millimeter wavelength range (**Fig. 1**), a microwave beam is directed at a beamsplitter, which splits the beam into two parts, A and B, by partial reflection. The A beam

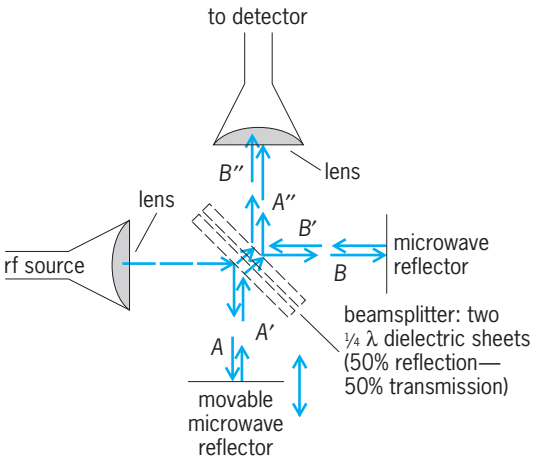


Fig. 1. Wavelength measurement by the Michelson interferometer used at millimeter wavelengths.

is reflected to a movable reflector and reflected again as A'. The beamsplitter transmits part of this as A''. The transmitted part B of the original beam is reflected by a fixed microwave reflector as B'. This is partially reflected by the beamsplitter as B''. The beams A'' and B'' combine to form standing waves, which are then detected. Movement of the movable reflector causes the position of the standing wave to move, which causes the detected signal to pass through successive points of maximum and minimum amplitude. The distance between points of successive maxima or minima is one-half wavelength. This distance may be determined from the motion of the movable reflector. *See* INTERFEROMETRY.

Standing-wave methods. A crude form of interferometer, used by early radio experimenters, does not require a beamsplitter. Instead it relies upon the standing-wave formed when a wave is reflected onto itself, as would occur, for example, with beams B and B' in the interferometer discussed above (Fig. 1). The maxima and minima of this standing wave can be observed by moving a suitable detector to various points in the field. Under ideal conditions the distance between successive minima or maxima is half the wavelength and can be measured with a measuring tape; but the detector and stray reflections can easily distort the observed pattern.

Lecher wires. In the Lecher-wire wavemeter the wave is made to travel down a transmission line. With this simple device, wavelength is measured by sliding a short circuit along the line and observing the cyclic variations of a power indicator. The distance between two successive absorption maxima or minima is half the wavelength ($\lambda/2$); thus by a simple length measurement the wavelength is measured directly.

Microwave wavemeters. These wavemeters make use of resonant coaxial-line sections or cavities as tuned elements. The two general types of microwave wavemeters are the absorption, or reaction, type and the transmission type. Wavemeters of low or medium selectivity are frequently used as coarse measuring devices to establish the general range of frequency of operation of a system before applying more refined and complex methods for accurate frequency checking. *See* CAVITY RESONATOR; COAXIAL CABLE.

The dimensions of the cavity determine the resonant frequency of a resonant-cavity microwave wavemeter (**Fig. 2**). A signal is fed in from either a coaxial line or waveguide, and energy is fed out to a suitable detector by a second coaxial line. The cavity is tuned by means of a micrometer-driven plunger, which may be calibrated in terms of wavelength. *See* MICROWAVE.

Tuned circuits. For wavelengths greater than a few meters, the dimensions of transmission lines and resonant cavities become inconveniently large. It is then more convenient to use resonant systems made from inductors and capacitors. With calibrated values of inductance L and capacitance C, it is possible to provide a scale calibrated in wavelength or frequency. For low-loss circuits the resonant frequency is $1/(2\pi\sqrt{LC})$. Absorption-type wavemeters (**Fig. 3**), often constructed with the principal inductance as a plug-in coil, are used for frequency or wavelength measurement up to frequencies of approximately 1000 MHz. *See* RESONANCE (ALTERNATING-CURRENT CIRCUITS).

Infrared and optical regions. The free-space wavelengths of monochromatic visible and infrared radiations can be derived directly from Eq. (1) if their corresponding frequencies are known. Air wavelengths can be determined by dividing the vacuum wavelength by the refractive index of air. The optical frequencies of a number of standards have been precisely determined by measurement of their frequency referenced to the cesium-133 primary frequency standard, using femtosecond combs. The most accurate standards published to date are based on transitions in single trapped ions and have an uncertainty within a factor of 3 or better of the cesium standard. These optical standards are used for the realization of the unit of length, but are also expected to become secondary representations of the second. *See* LENGTH; LIGHT; PHYSICAL MEASUREMENT; WAVELENGTH STANDARDS; FREQUENCY MEASUREMENT.

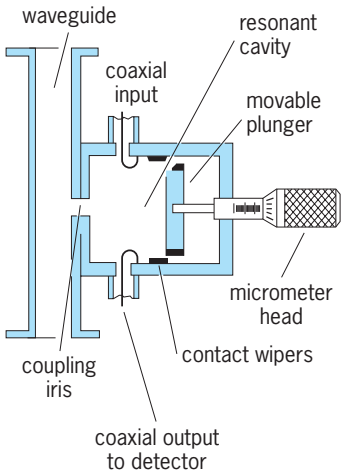


Fig. 2. Typical construction of a resonant-cavity wavemeter providing either coaxial or waveguide inputs.

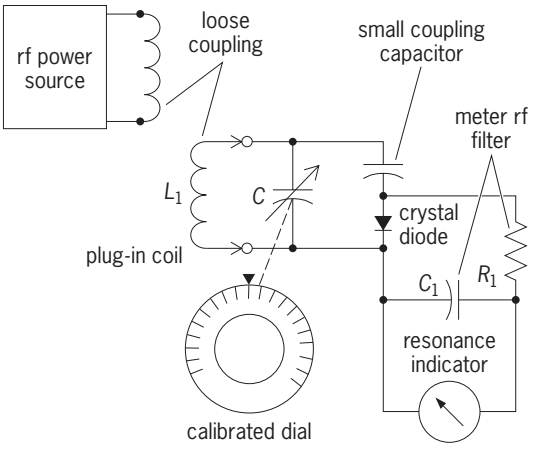


Fig. 3. Schematic diagram of inductance-capacitance type of absorption wavemeter (for frequencies between approximately 50 kHz and 1000 MHz).

Dispersion methods. Wavelength values to an accuracy of 1 part in 10^5 can be determined with a spectrometer, spectrograph, or monochromator, in which a prism or diffraction grating is used as a dispersive element. Each wavelength forms a line image of the entrance slit at a particular angle. An unknown wavelength can be determined by interpolation with the pattern formed by a lamp emitting the tabulated characteristic wavelengths of a particular element. Atomic reference data for this purpose are published, for example, on the Internet (see, for example, http://physics.nist.gov/PhysRefData/ASD/index.html). Care must be taken in using tabulated wavelength values because the values may be either for vacuum or for the refractive index of standard air. (Air wavelengths are smaller than vacuum values by about 1 part in 3000.) *See* DIFFRACTION GRATING; OPTICAL PRISM; SPECTROSCOPY.

Use of interferometers. The most precise wavelength measurements use an interferometer to compare the unknown wavelength λ_1 with a standard wavelength λ_2. Usually either the two-beam Michelson form or the multiple-beam Fabry-Perot form of interferometer is used. The general equation, applicable to both forms, is (2), in which θ is the angle incidence, t is

$$(m_1 + f_1)\lambda_1 = (m_2 + f_2)\lambda_2 = 2nt\cos\theta \qquad (2)$$

the real or virtual separation of the reflectors, and n is the refractive index of the medium between the reflectors. Thus at any arbitrary reference point in the interference pattern, Eq. (3) holds, where m_1 and

$$\lambda_1 = \frac{\lambda_2(m_2 + f_2)}{m_1 + f_1}$$

m_2 are integers, usually called the orders of interference, and f_1 and f_2 are fractions. Maximum transmission of light of wavelength λ occurs when its corresponding fractional order f is zero. Interferometer measurements to an accuracy of a few megahertz are necessary to determine the mode order number of femtosecond comb systems used to measure optical frequencies.

Several forms of interferometric "wavemeter" are commercially available for use with laser radiations. One of their principal uses is to measure the wavelength emitted by a tunable laser, for example a titanium-sapphire or diode laser, to better than 1 part in 10^6, so that it may be tuned into coincidence with a desired spectral transition. One form of wavemeter has a retroreflector that moves to and fro along a track. This reflector forms part of a two-beam interferometer so that sinusoidal intensity signals are generated. The signals for the unknown and standard radiations are counted electronically, giving totals that correspond to the order numbers m_1 and m_2. Another form of laser wavemeter, suitable for use with pulsed as well as continuous lasers, uses the wedged Fizeau interferometer with a detector array to measure the spacing and position of the interference pattern. *See* LASER; LASER SPECTROSCOPY.

Fourier transform method. When a number of wavelengths are mixed together in the input to a moving-carriage two-beam interferometer, the output signal is the summation of the many separate sine-wave signals having different periods. A Fourier analysis of this composite signal enables the separate wavelengths to be identified. This Fourier transform method is particularly useful for the measurement of complex spectra in the infrared. *See* FOURIER SERIES AND TRANSFORMS; INFRARED SPECTROSCOPY.

W. R. C. Rowley; Geoffrey P. Barwood

Bibliography. G. H. Bryant, *Principles of Microwave Measurements*, Institution of Electrical Engineers, 1993; P. Hariharan, *Optical Interferometry*, Academic Press, 2d ed., 2003; J. M. Hollas, *Modern Spectroscopy*, 4th ed., Wiley, 2004; T. S. Laverghetta, *Modern Microwave Measurements and Techniques*, rev. ed., Artech House, 1989; P. J. Mohr and B. N. Taylor, CODATA recommended values of the fundamental physical constants, *Rev. Mod. Phys.*, 77:1–107, 2005; T. J. Quinn, *Mise en pratique* of the definition of the metre (2001), *Metrologia*, 40:103–133, 2003; B. N. Taylor and W. D. Phillips, *Precision Measurements and Fundamental Constants II*, Nat. Bur. Stand. Spec. Publ. 617, 1984.

Wavelets

The elementary building blocks in a mathematical tool for analyzing functions. The functions to be analyzed can be very diverse; examples are solutions of a differential equation, and one- and two-dimensional signals. The tool itself, the wavelet transform, is the result of a synthesis of ideas from many different fields, ranging from pure mathematics to quantum physics and electrical engineering.

In many practical applications, it is desirable to extract frequency information from a signal— in particular, which frequencies are present and their respective importance. An example is the decomposition into spectral lines in spectroscopy. The tool that is generally used to achieve this is the Fourier transform. Many applications, however, concern so-called nonstationary signals, in which the makeup

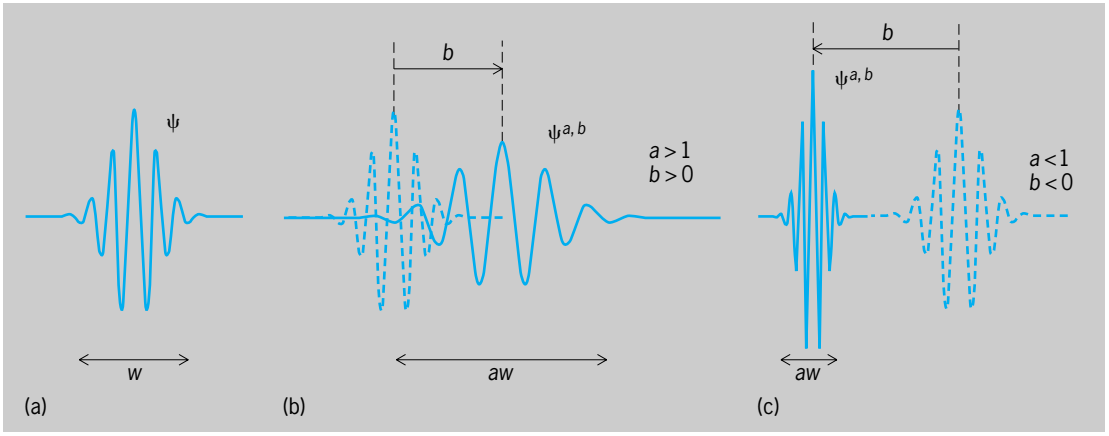


Fig. 1. Members of a wavelet family: an original and two shifted and dilated copies. (*a*) Original wavelet ψ with width *w*. (*b*) Large-scale wavelet $\psi^{a,b}$, with width *aw* that is larger than the original width *w*. (*c*) Fine-scale wavelet $\psi^{a,b}$ with narrow width.

of the different frequency components is constantly shifting. An example is music, where this shifting nature has been recognized for centuries by the standard notation, which tells a musician which note (frequency information) to play when and how long (time information). For signals of this nature, a time-frequency representation is needed. *See* FOURIER SERIES AND TRANSFORMS.

There exist many different mathematical tools leading to a time-frequency representation of a given signal $f(t)$, each with its own strengths and weaknesses. The wavelet transform is such a time-frequency analysis tool. Its particular strength lies in its ability to deal well with transient high-frequency phenomena, such as sudden peaks or discontinuities, as well as with the smoother portions of the signal. (An example is a crack in the sound from a damaged record, or the attack at the start of a music note.) The wavelet transform is less well adapted to harmonically oscillating parts in the signal, for which Fourier-type methods are more indicated.

Wavelet families. A family of wavelets is derived from one function by dilation and translation. The original function ψ is typically smooth and well concentrated; the whole family is then given by Eq. (1),

$$\psi^{a,b}(x) = |a|^{-1/2}\psi\left(\frac{x-b}{a}\right) \tag{1}$$

where a and b are real numbers and a is greater than 0. If the original wavelet ψ is centered on 0 and has width w, then $\psi^{a,b}$ is centered on b and has width aw (**Fig. 1**). The parameter b indicates the time location (if the variable x stands for time), while a gives the scale. The scale parameter is related to frequency: If the original wavelet has approximately k oscillations over its width w, corresponding (coarsely speaking) to a frequency $\omega_0 = k/w$, then $\psi^{a,b}$ will have the same number of oscillations for a width aw, corresponding to frequency $k/aw = \omega_0/a$. Large a corresponds therefore to long wavelets with low frequency, and very small a to narrow, high-frequency wavelets. In practice the range of scales can bridge several orders of magnitude.

Signal analysis. Once a family of wavelets is given, it can be used to analyze signals. Wavelets can be used to characterize or analyze a signal $f(x)$ in two different ways. The first is to correlate the signal f with all the wavelets in the family, that is, to compute the wavelet transform, given by Eq. (2). The second is

$$(Wf)(a,b) = \int f(x)\psi^{a,b}(x)\,dx \tag{2}$$

to try to write f as a superposition of wavelets, with weights which are called the wavelet coefficients. In the case where both the scale and translation parameters are allowed to vary continuously, the superposition is an integral, given by Eq. (3). If the wavelet

$$f(x) = \int_0^\infty \int_{-\infty}^\infty (wf)(a,b)\psi^{a,b}(x)\frac{da\,db}{a^2} \tag{3}$$

ψ is a real function with good localization and zero integral, then the weights in the superposition can be chosen to satisfy Eq. (4), where C_ψ is a constant

$$(wf)(a,b) = C_\psi(Wf)(a,b) \tag{4}$$

depending only on ψ (not on f). This shows that the two approaches, characterizing a function f by means of its inner products with $\psi^{a,b}$, as in Eq. (2), or decomposing it, as in Eq. (3), into a superposition of $\psi^{a,b}$, amount to the same thing.

Discrete parameters. In Eqs. (2) and (3), the scale and localization parameters a and b vary continuously; this is the continuous wavelet transform, a useful analysis tool, but very redundant. In another

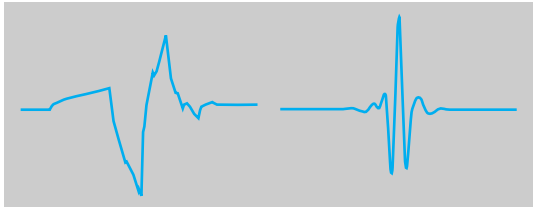


Fig. 2. Two examples of wavelets which generate an orthonormal basis, and which are associated with fast algorithms.

version of the wavelet transform, a and b are restricted to a discrete lattice of values, such as $a = 2^j$, $b = 2^j k$, where j and k are integers. In this case the superposition integral in Eq. (3) is replaced by the sum in Eq. (5), where the factors on the right-hand side are given by Eqs. (6). In order for Eq. (5) to hold,

$$f(x) = \sum_{j,k} \langle f, \psi_{j,k} \rangle \psi_{j,k}(x) \tag{5}$$

$$\psi_{j,k}(x) = 2^{-j/2} \psi(2^{-j}x - k) \tag{6a}$$

$$\langle f, \psi_{j,k} \rangle = \int f(x) \psi_{j,k}(x)\, dx \tag{6b}$$

many more restrictions on ψ have to be imposed than in the continuous case. However, the wavelets $\psi_{j,k}$ are orthonormal, and there is no redundancy in the reproduction of f by the sequence $\langle f, \psi_{j,k} \rangle$ [**Fig. 2**]. *See* NONRELATIVISTIC QUANTUM THEORY.

Data compression. Smooth orthonormal wavelet bases are interesting because their smoothness makes them suitable for the decomposition of smooth functions, whereas the different scales in the wavelet family make it possible to fine-tune the decomposition locally near singularities or transients. The combination of smoothness and scaling in orthonormal wavelet bases implies that for many functions or signals f of interest, large numbers of fine-scale coefficients are negligibly small, because they are localized in smooth, uneventful parts of f, which are already well captured by the wider-scale wavelets. Because these tiny coefficients can be neglected without appreciable distortion in the reconstruction of f, data compression is possible. *See* DATA COMPRESSION.

Discrete variables. In the formulas above, x is a continuous variable. Similar formulas can be written if x varies discretely, on a lattice; the integrals then become sums. Most practical implementations use such a discrete version. In this case the orthonormal wavelet transform (and its generalizations) can be implemented via fast algorithms [with complexity of order $O(N)$ for data with N samples] that are related to subband filtering in electrical engineering.

Applications. Applications of wavelets include various forms of data compression (such as for images and fingerprints), data analysis (nuclear magnetic resonance, radar, seismograms, and sound), and numerical analysis (fast solvers for partial differential equations). *See* DIFFERENTIAL EQUATION; INTEGRAL TRANSFORM; NUMERICAL ANALYSIS. Ingrid Daubechies

Bibliography. I. Daubechies, *Ten Lectures on Wavelets*, 1992; Y. Meyer, *Ondelettes*, 1990, transl. as *Wavelets*, 1993; Y. Meyer, *Wavelets: Algorithms and Applications*, 1993.

Wavellite

A hydrated phosphate of aluminum mineral with composition $Al_3(OH)_3(PO_4)_2 \cdot 5H_2O$, in which small amounts of fluorine and iron may substitute for the hydroxyl group (OH) and aluminum (Al), respectively. Wavellite crystallizes in the orthorhombic system. The crystals are stout to long prismatic, but are rare. Wavellite commonly occurs as globular aggregates of fibrous structure (see **illus.**) and as encrusting and stalactitic masses. Wavellite crystals range in color from colorless and white to different shades of blue, green, yellow, brown, and black.

Wavellite in globular aggregates, found in Devonshire, England. (*Specimen from Department of Geology, Bryn Mawr College*)

Wavellite is a widespread secondary mineral occurring in small amounts in crevices of low-grade metamorphic aluminous rocks, in limonite, and in phosphate rock deposits. Found at many places in Europe and North America, it is also abundant in tin veins at Llallagua, Bolivia. *See* PHOSPHATE MINERALS.

Wayne R. Lowell

Wax, animal and vegetable

Any of the substances containing esters of higher fatty acids and long-chain monohydric alcohols. From a practical standpoint, this definition is inadequate because it allows liquids such as sperm whale oil and jojoba oil to be called waxes, and it fails to indicate the complexity of waxes. While waxes do contain wax esters, they are seldom if ever pure. They are usually mixtures that may contain high-molecular-weight acids, alcohols, esters, ketones, hydrocarbons, sterols, diesters, hydroxyacids, and so forth, as well as the wax esters. *See* ESTER.

Properties. Practical wax formulators use physical, rather than chemical, properties to define waxes. Waxes must: be a solid at 20°C (68°F); be crystalline; melt above 40°C (104°F) without decomposition,

Physical properties of selected waxes

Substance	Melting point, °C (°F)	Acid value*	Saponification value†	Iodine value‡	Acetyl value§
Beeswax	63 (145)	19	92	10	15
Candelilla	67 (153)	15	55	33	
Carnauba	84 (183)	5	83	10	55
Spermaceti	45 (113)	1	120	2	
Wool wax	40 (104)	20	100	30	

*mg KOH/g of wax to neutralize free acid.
†mg KOH/g of wax to saponify esters.
‡cg I_2/g of wax absorbed; indicates amount of unsaturation.
§mg KOH to neutralize acetic acid from saponifying 1 g acetylated wax; indicates amount of free hydroxyl.

have relatively low viscosity above the melting point, have consistency and solubility properties that are strongly dependent upon temperature; and be capable of being polished under slight pressure. Typical properties of several waxes are shown in the **table**.

Sources. Wax sources are abundant; therefore it is the quantity and ease of recovery that generally determine whether a wax becomes commercially valuable. Waxes have been isolated from the outer layers of bacteria, the roots, stems, leaves, fruit, and flowers of plants, the exudates of insects, the skin and hair of some animals, and the bodies of certain marine and land animals.

Important waxes. Carnauba wax is extracted from an exudate on the leaves of the carnauba palm (*Copernicia prunifera*). The only significant production is in northeastern Brazil. Leaves are cut from the tree and allowed to dry. During drying the leaf shrinks, and the wax is no longer adherent. Machines beat the leaves to free the wax, and so produce a mixture of crude wax and leaf particles which is melted and stained to isolate the crude product. Carnauba is valuable because it has a relatively high melting point. It buffs to a high luster without showing the polish lines. *See* CARNAUBA WAX.

Candelilla wax is obtained from a coating on the stem of *Euphorbia antisyphilitica*, a leafless desert shrub. It is fairly common in the Big Bend country of Texas, but the commercial product comes from the states of Chihuahua and Coahuila in Mexico. The shrubs are pulled up when they are about 1 m tall, bundled, and heated in water. The wax rises to the surface, and is skimmed off. Candelilla wax contains a tacky resin, as well as large amounts of hydrocarbon. It has strong adhesive properties, and the resin imparts antislip characteristics to formulated waxes.

Beeswax is an exudate of the honeybee. Most of the commercial beeswax comes from Africa. Hollow logs hung from trees are used by wild bees as hives. When the logs are presumed to be full, they are removed, and the wax is recovered by melting. Beeswax is an important ingredient in many cosmetic formulations. Considerable quantities are used in candles, with lesser amounts going into polishes, modeling, and pattern making. Beeswax is a good solvent for other waxes; it is a plasticizer, and promotes adhesion. *See* BEEKEEPING.

Wool wax, obtained from sheep wool, is refined to produce lanolin, which is very hydrophilic compared to other waxes. It is widely used in cosmetics and pharmaceutical creams and salves. Its property of softening dry skin without affecting natural moisture transpiration is probably due to the large amounts of sterols and alcohols present. *See* LANOLIN.

Other waxes. Many other waxes have been of commercial importance from time to time, but their availability or cost has forced them from the market. Sperm oil and spermaceti are extracted from the blubber and head cavities of the sperm whale (*Physeter macrocephalus*). Spermaceti has properties that are useful in cosmetics, but since the sperm whale is an endangered species, it cannot be hunted. Instead, the spermaceti market is satisfied by esters derived from fatty acids and synthetic fatty alcohols. Ouricouri wax is similar to carnauba. It is a coating on palm leaves, but unlike carnauba the wax must be scraped since it does not flake free from the leaf. The decrease in sales of ouricouri is probably due to high labor costs. Retamo wax is similar to candelilla. It comes from a desert shrub in northwestern Argentina. Government controls raised the price so high that it is no longer on the market. Sugarcane wax is extracted from the filtrate in refining cane sugar.

Analysis. Until the advent of gas-liquid chromatography, it was virtually impossible to analyze waxes, but the important characteristics of a wax are still based on physical properties or empirical chemical reactions. *See* FAT AND OIL.

Howard M. Hickman

Wax, petroleum

A substance produced primarily from the dewaxing of lubricating-oil fractions of petroleum. It may be of either the crystalline or microcrystalline type. The crystalline wax is produced from distillate lubricating fractions, whereas the microcrystalline wax is obtained from the residual lubricating fractions of the crude oil. The melting-point range for refined crystalline waxes is 120–150°F (48–65°C) while the petrolatum or microcrystalline waxes have melting points in the range of 150–175°F (65–79°C).

Petroleum waxes constitute approximately 90% of all wax used in industry today. The other 10% comprises vegetable and animal waxes used primarily as specialty waxes. *See* WAX, ANIMAL AND VEGETABLE.

Petroleum wax and petrolatum production in the United States has varied depending upon the economic demand.

Uses. Petroleum wax has a wide variety of uses. It is used to coat paper products, to blend with other waxes for the manufacture of candles, in the manufacture of electrical equipment and many polishes for home and industry, and as a source material for oxidized products which are being more widely used in industry. The softer waxes, such as petroleum jelly, after proper purification, are being used as medicinal products.

The petrolata have many uses, principally as components of blended waxes used for coating paper, electrical insulation, and water- and rustproofing materials, as components of crayons and printing inks, and in the manufacture of specialties. *See* PETROLATUM.

In addition to melting point, other properties are important in establishing the quality and end use of waxes. These are tensile strength, hardness, stability, odor and taste, oil content, blocking and gloss characteristics, scuff resistance, and sealing strength.

Purification of crude wax. The slack wax obtained from the primary dewaxing operation on the lubricating fraction must be purified further in most cases before marketing. The first operation in this purification is the removal of all or most of the oil associated with the wax. The deoiling operation on the wax is conducted by either the sweating operation or the solvent deoiling operation.

The old and widely used sweating process is, in reality, a fractional separation of crude wax (slack wax) obtained from a dewaxing process. The separation by melting points is accomplished in large pans or sweaters contained in a well-insulated building which can be heated uniformly at a slow rate to effect the melting-point separations. To accomplish this, molten slack wax is placed in the sweater pans and is cooled until a solid cake is formed. The sweater pans are then gradually warmed and the waxes are separated—the lowest-melting-point fractions being separated as liquid fractions or cuts. As the temperature of the sweater is increased, the higher-melting-point waxes are removed. The cuts or fractions are then blended to provide the desired range of melting point. The melting points of the fractions vary from 110 to approximately 160°F (43 to 71°C). These fractions or blends are usually treated with sulfuric acid followed by clay filtration and deodorizing for the removal of impurities. The finished wax is molded into various shapes for commercial sale and distribution to industry. The commercial grades of paraffin wax are known as crude scale and fully refined waxes, and they vary in melting-point range.

The solvent deoiling and fractionation process is a relatively new method of separating wax into its various melting-point fractions. It rapidly is replacing the sweating process because of its flexibility and wider range of application. This process is quite similar to the solvent dewaxing process. The slack wax to be deoiled or fractionated is dissolved in a suitable solvent and the mixture chilled to precipitate the desired melting-point grade of wax. The wax is separated by filtration from the solvent solution which retains the more soluble waxes and oil. Further chilling and filtration steps separate additional wax fractions. Additional treatment with acid and clay results in high-quality waxes.

The petrolatum type of wax cannot be made by the sweating processes. These waxes are made directly from residual lubricating oil from which they are separated either by the solvent dewaxing process or by the centrifuge dewaxing method. Fractional separation of petrolatum into various melting points can be accomplished by the solvent deoiling process. One of the first methods of obtaining petrolatum was through the recovery of the wax bottoms that had settled out of wax-bearing crude oils during storage. *See* DEWAXING OF PETROLEUM; PETROLEUM PROCESSING AND REFINING; PETROLEUM PRODUCTS.

Wayne E. Kuhn

Weak nuclear interactions

Fundamental interactions of nature that play a significant role in elementary-particle and nuclear physics, and are distinguished from other such interactions by special properties such as participation of all the fundamental fermions and failure to conserve parity and to respect particle–antiparticle symmetry. According to present understanding, the four fundamental forces of nature are the gravitational, the electromagnetic, the strong, and the weak. The weak force has short range (less than 10^{-17} m) and at low energies is very feeble compared to the strong and electromagnetic forces, but it can be distinguished from the latter by its special character. For example, all of matter (with the possible exception of dark matter) appears to consist of certain basic constituents, the quarks and leptons, collectively called the fundamental fermions (**Table 1**). While only quarks participate in strong interactions, and only the quarks and the charged leptons ($e^{\pm}$, $\mu^{\pm}$, $\tau^{\pm}$) participate in electromagnetic interactions, all of the fundamental fermions, including neutrinos, engage in weak interactions. Moreover, while the strong and electromagnetic interactions respect spatial inversion symme-

TABLE 1. Fundamental fermions*

Leptons	Family		
Electric charge	1	2	3
0	ν_e	ν_μ	ν_τ
$-e$	e^-	μ^-	τ^-
Quarks†	Family		
Electric charge	1	2	3
$2e/3$	u	c	t
$-e/3$	d	s	b

*Corresponding to each fermion, there is an antifermion with the same mass and opposite electric charge.
†Quarks are confined in mesons and baryons and do not appear as free particles.

TABLE 2. Classification of representative weak interactions*

Interaction	Charged	Neutral
Leptonic		
	$\mu \to e\nu\nu$	$\nu_\mu e \to \nu_\mu e$
	$\nu_e e \to \nu_e e$	$\nu_e e \to \nu_e e$
	$\tau \to l\nu\nu$	$e^+e^- \to \nu_\mu\overline{\nu}_\mu$
Semileptonic		
Meson	$\pi^+ \to \mu\nu,e\nu,\pi^0 e\nu$	
	$K^+ \to \mu\nu,e\nu,\pi^0 e\nu$	
	$K_L^0 \to \pi^\pm\mu\nu,\pi^\pm e\nu$	
Baryon	$B \to B'l\nu$	$\nu N \to \nu N,\nu N\pi,\nu NX$
	$\mu^- B \to B'\nu$	$\nu + D \to n + p + \nu$
	$\nu B \to B'l$	$eN \to eN,eX$
Nonleptonic		
Meson	$K \to \pi\pi$	
	$K \to 3\pi$	
	$D \to KK,K\pi,K2\pi,...$	
	$B \to D\pi,DK,...$	
Baryon	$\Lambda \to N\pi$	$NN \to NN$
	$\Sigma \to N\pi$	
	$\Xi \to N\pi$	

*N refers to a nucleon, B refers to a baryon, l refers to a charged lepton, and X refers to highly excited hadronic matter in deep inelastic collisions. Some reactions (for example, $\nu_e e \to \nu_e e$) can occur by neutral as well as charged weak interactions. The reaction $eN \to eN$ occurs by electromagnetic as well as weak interactions. The reaction $NN \to NN$ occurs by strong and electromagnetic as well as weak interactions.

try (parity) and are also particle–antiparticle (charge conjugation) symmetric, the weak interaction violates these two symmetries. *See* FUNDAMENTAL INTERACTIONS; LEPTON; PARITY (QUANTUM MECHANICS); QUARKS; SYMMETRY LAWS (PHYSICS).

Weak interactions are classified as charged or neutral, depending on whether or not a particle participating in a weak reaction suffers a change of electric charge of one electronic unit (**Table 2**). For example, the neutrino–nucleon scattering reaction (1) is

$$\nu_\mu + n \to \mu^- + p \tag{1}$$

charged because the neutrino ν_μ, which has zero charge, transforms into a negative muon μ^-, while simultaneously a neutron n (zero charge) becomes a proton p (positive charge). Actually the valence quark composition of a neutron is two d quarks and a u quark, while that of a proton is two u quarks and a d quark. Thus in reaction (1) a d quark transforms to a u quark. The weak scattering reaction (2) is neutral

$$\nu_\mu + p \to \nu_\mu + p \tag{2}$$

because the final neutrino and proton retain their initial charges. Observed charged weak interactions include nuclear beta decay and electron capture, muon capture on nuclei, decays of the μ and τ leptons and the $\pi^\pm$ and K mesons, the slow hyperon decays, decays of c- and b-quarked mesons and baryons, and the decay of the top quark. Charged neutrino–nucleon and neutrino–lepton scattering reactions, and the phenomenon of neutrino oscillations, are additional manifestations of the charged weak interaction. Neutral weak interactions, which were first observed in 1973, include but are not limited to neutral neutrino–nucleon and neutrino–lepton scattering as well as the weak electron–nucleon reaction (3), which can also occur by electromagnetic

$$e^- + N \to e^- + N \tag{3}$$

interaction. *See* BARYON; ELEMENTARY PARTICLE; HYPERON; MESON; NEUTRAL CURRENTS.

Early history. Study of weak interactions began with the discovery of radioactivity by H. Becquerel in 1896 and the recognition shortly thereafter that in one form of radioactivity the decaying nucleus emits "beta rays" (electrons). Thus nuclear beta decay was the first known weak process. In 1914 J. Chadwick observed that the electrons in beta decay of a given isotopic sample are emitted with a continuous spectrum of energies. This result and subsequent observations in the 1920s led to a crisis, because the energy available to the electron in beta decay is essentially the difference in rest energies of the initial and final nuclei, which is a fixed quantity. Thus, it appeared that the principle of energy conservation was violated in beta decay. In order to rescue that principle, W. Pauli proposed in 1930 and again in 1933 that a neutral particle of small or vanishing rest mass and half-integral spin (later called the neutrino) is emitted along with the electron in nuclear beta decay, that it shares the available fixed energy with the electron (leading to the continous electron energy spectrum), and that it escapes observation because of its feeble interactions with surrounding matter. *See* BETA PARTICLES; NEUTRINO; RADIOACTIVITY.

Fermi's theory. In 1934 E. Fermi proposed a theory of beta decay based on Pauli's neutrino hypothesis. In Fermi's theory, beta decay is described by the coupling of two charged weak "current densities" at one common space-time point. For example, in neutron beta decay, reaction (4), one of these current densi-

$$n \to p + e^- + \overline{\nu}_e \tag{4}$$

ties describes the neutron-to-proton transformation, while the other accounts for the simultaneous creation of the electron and antineutrino. Each current density behaves as a four-vector (V) under Lorentz transformations. The strength of the coupling is characterized by the "Fermi coupling constant" G_F, a quantity that must be determined by experiment and is found to be given by Eqs. (5) or in units where $\hbar = c = 1$ (where $\hbar$ is Planck's constant divided by 2π and c is the speed of light) by Eq. (6).

$$G_F = 1.43584 \pm 0.00001 \times 10^{-49}\text{erg} \cdot \text{cm}^3 = 1.43584 \pm 0.00001 \times 10^{-62}\text{J} \cdot \text{m}^3 \tag{5}$$

$$G_F = 1.16637 \pm 0.00001 \times 10^{-5}\text{GeV}^{-2} \tag{6}$$

See RELATIVITY.

Fermi's theory gave a good phenomenological account of many aspects of nuclear beta decay, especially when generalized by G. Gamow and E. Teller in 1936. It also anticipated a number of future developments, for in the 25 years following Fermi's proposal, new weak processes were found in addition to nuclear beta decay, with the discovery of new

elementary particles and elucidation of their decay schemes. Gradually it became clear that these bear many similarities to nuclear beta decay and appear to be different manifestations of a universal "Fermi" interaction. However, it was recognized quite early (by W. Heisenberg in 1936, among others) that the Fermi theory, while undoubtedly very useful, cannot be fundamental, because it leads to a failure of unitarity (nonconservation of probability) when it is applied to high-energy processes such as neutrino–electron scattering. A related and serious shortcoming is that the Fermi theory is not renormalizable: It contains incurable divergences that occur in the calculation of higher-order corrections. *See* RENORMALIZATION.

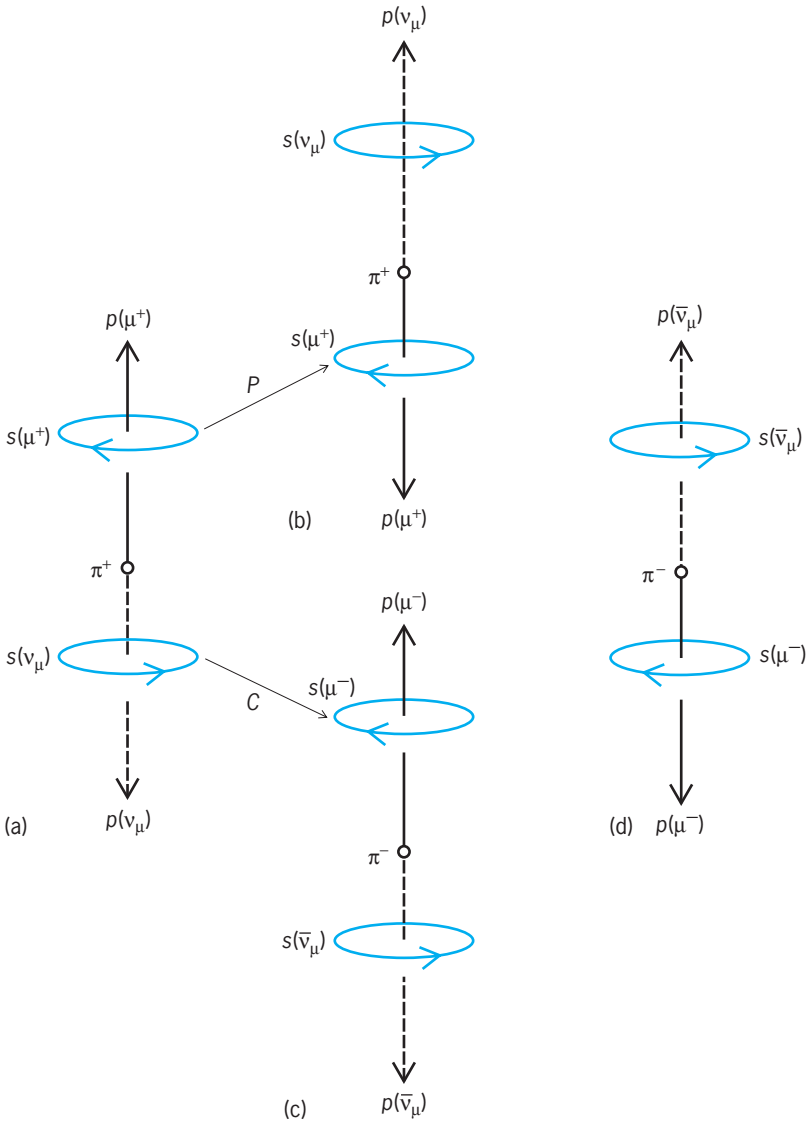


Fig. 1. Weak decay of the pion, illustrating parity (*P*) and charge conjugation (*C*) violation in weak interactions. (*a*) The decay $\pi^+ \rightarrow \mu^+ + \nu_\mu$. Experiment shows that μ^+ spin [here labeled $s(\mu^+)$] is opposed to its momentum $p(\mu^+)$. In the pion rest frame, the neutrino momentum $p(\nu_\mu)$ must be equal and opposite to $p(\mu^+)$. Also, since the pion has zero spin angular momentum, the neutrino spin $s(\nu_\mu)$ must be opposed to $s(\mu^+)$. (*b*) The result of a parity transformation on *a*. The momenta $p(\mu^+)$ and $p(\nu_\mu)$ are reversed, but the spins $s(\mu^+)$ and $s(\nu_\mu)$ remain unchanged. Here $p(\mu^+)$ and $s(\mu^+)$ are parallel, which is never observed. (*c*) The result of a charge conjugation transformation on *a*. Here the spins and momenta are left unchanged, but $\pi^+ \rightarrow \pi^-$, $\mu^+ \rightarrow \mu^-$, $\nu_\mu \rightarrow \overline{\nu}_\mu$. The arrangement shown here is never observed; experiment shows that in π^- decay the μ^- spin and momentum are always parallel. (*d*) The result of a *CP* transformation on *a*. This figure correctly represents the decay $\pi^- \rightarrow \mu^- + \overline{\nu}_\mu$

Parity nonconservation. In the 1950s the accumulation of data on the weak decays of what is now known as K^+ mesons led to a peculiarity called the tau–theta ($\tau - \theta$) puzzle. A search for the solution of this puzzle induced T. D. Lee and C. N. Yang in 1956 to propose a dramatic, important, and far-reaching hypothesis: Parity is violated in weak interactions. Their ideas were almost immediately subjected to experimental test and quickly vindicated. An example of parity violation appears in the weak decay of the charged pion: $\pi^+ \rightarrow \mu^+ + \nu_\mu$ (**Fig. 1**). In observed $\pi^+ \rightarrow \mu^+ + \nu_\mu$ decay (Fig. 1*a*), the μ^+ spin is found experimentally to be opposite to its linear momentum: helicity $h(\mu^+) = -1$. Since the π^+ spin is zero and angular momentum as well as linear momentum is conserved, the neutrino must be emitted in the opposite direction to that of the muon, and its helicity must also be $h(\nu_\mu) = -1$. A parity (P) transformation is equivalent to a mirror reflection in the plane containing the pion and perpendicular to the lepton linear momenta. This transformation reverses the muon and neutrino linear momenta but leaves their spins invariant. Thus under P, $h(\mu^+)$ and $h(\nu_\mu)$ are reversed, but this outcome (Fig. 1*b*) is never observed. Since parity invariance requires that a physical process and its mirror image occur with the same probability, one has here a "maximal" parity violation. One can also make a charge conjugation transformation on the particles of Fig. 1*a*. This changes the particles into the corresponding antiparticles but leaves the spins and linear momenta, and hence the helicities, invariant (Fig. 1*c*). However, the process of Fig. 1*c* is never observed, for experiment shows that in π^- decay, $h(\mu^-) = +1$. Thus, charge conjugation symmetry (C) is also violated maximally. On the other hand, a combined charge conjugation and parity (CP) transformation on $\pi^+ \rightarrow \mu^+ + \nu_\mu$ results in observed $\pi^- \rightarrow \mu^- + \overline{\nu}_\mu$ decay (Fig. 1*d*). *See* HELICITY (QUANTUM MECHANICS).

V-A law. The discovery of P and C violation inspired intense experimental and theoretical activity that led to the V-A law, proposed by E. Sudarshan and R. Marshak, and independently by R. Feynman and M. Gell-Mann, in 1958. In this formulation Fermi's theory was generalized in two respects: first, parity violation was accounted for by including an axial vector (A) as well as a vector (V) portion in each weak current density; and second, the Fermi-type couplings were generalized to describe not only beta decay but also muon decay, decays of charged pions, and so forth. This scheme was thus able to give an excellent account of the observed features of low-energy charged weak processes, including the newly discovered parity-violating effects, but since it was nothing more than an enhanced version of Fermi's theory, it suffered from the same defects: It is not renormalizable and leads to difficulties (breakdown of unitarity) at high energies. A radical change in point of view was thus necessary.

Emergence of the standard model. The path to a better theory starts with consideration of the possibility that a weak interaction does not occur at a single point in space-time as in Fermi-type

theories, but proceeds instead through the exchange of an intermediate boson. This idea arises naturally by analogy with the original meson theory of nuclear forces (H. Yukawa, 1935) and with quantum electrodynamics (QED), where the Coulomb force between two charged particles occurs by exchange of a photon, the spin-1 quantum of the electromagnetic field (**Fig. 2***a*). The Coulomb force is long-range (it varies as $1/r^2$), which means that the photon mass is zero. However, experiment shows that the weak force, like the nuclear force, has very short range. Therefore, if intermediate bosons (also spin-1 quanta) mediate the weak interaction, they must be massive. Also, there must be at least three such bosons: a charged W^- and its charge conjugate W^+ to transmit the charged weak interaction (Fig. 2*b*), and a neutral boson Z^0 for the neutral weak interaction (Fig. 2*c*). Thus, the problem may be stated as follows: How can one construct a renormalizable theory of weak interactions in which massive vector bosons, both charged and neutral, are exchanged between fundamental fermions?

The solution was indeed found in a theory uniting weak and electromagnetic interactions, and it was based on a combination of subtle ideas. First is the notion that the theory must be invariant under local gauge transformations, which are phase transformations of a quantum field that can vary in an arbitrary manner from one space–time point to another. Quantum electrodynamics is such a theory. Here the gauge symmetry group is U(1), the group of all 1×1 unitary matrices. The quanta of the electromagnetic field that emerge from these considerations are photons, which carry no charge. In 1954, Yang and R. Mills considered an analogous theory that employs instead the gauge group SU(2), the group of all 2×2 unitary matrices with determinant equal to unity. In this way Yang and Mills obtained a renormalizable theory similar to quantum electrodynamics but containing a triplet of spin-1 quanta with "charges" $+g$, $-g$, and zero, where g is the coupling constant in their theory. These quanta seem to be attractive candidates for the weak vector bosons. However, the Yang–Mills quanta are massless, while weak vector bosons must necessarily be massive. *See* GAUGE THEORY.

The second important idea, which overcomes this obstacle, is spontaneous symmetry breaking. This means that one has a quantum field theory possessing a certain symmetry not shared by the ground state of the system. It was shown by J. Goldstone in 1961 that when this situation occurs in a non-gauge field theory, a massless spin-0 excitation—the so-called Goldstone boson—emerges, corresponding to each degree of freedom in which the symmetry is broken. There is no experimental evidence for such bosons, so it would appear that spontaneous symmetry breaking does not occur. However, the proof of Goldstone's theorem is based on two assumptions, and in a gauge theory (for example, electrodynamics or the Yang-Mills theory) these assumptions cannot both be valid. Thus gauge theories evade the Goldstone theorem, as was first noted by P. Higgs in 1964. A most extraordinary result of these arguments is that the difficulties associated with massless

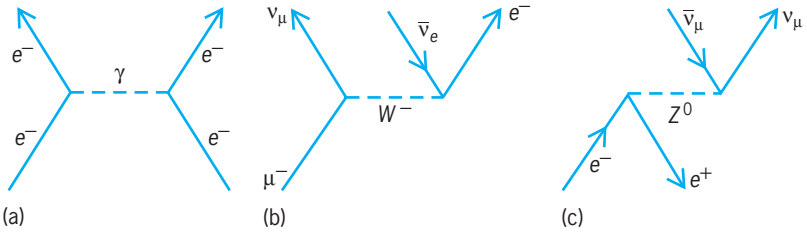


Fig. 2. Schematic diagrams of representative intermediate boson exchange processes between fermions. In these diagrams time increases upward. (*a*) Electron–electron scattering by photon exchange (electromagnetic interaction). (*b*) Muon decay ($\mu^- \rightarrow e^- + \bar{\nu}_e + \nu_\mu$) by W^- exchange (charged weak interaction). The corresponding charge conjugate process ($\mu^+ \rightarrow e^+ + \nu_e + \bar{\nu}_\mu$) by W^+ exchange occurs with the same total probability. (c) Neutrino pair production ($e^- + e^+ \rightarrow \nu_\mu + \bar{\nu}_\mu$) by Z^0 exchange (neutral weak interaction).

vector bosons and Goldstone bosons can be made to neutralize one another: By a suitable transformation, the Goldstone bosons disappear and simultaneously the vector bosons acquire mass. An inevitable consequence, however, is the appearance of at least one type of massive scalar particle, called a Higgs boson. Although Higgs bosons are required by the theory, no definite prescription can be given in this theory for their mass, which can be anywhere between about 7 GeV/c^2 and about 1000 GeV/c^2 a priori. Employing these ideas and others contributed earlier by S. Glashow, S. Weinberg (1967) and independently A. Salam (1968) constructed a successful gauge theory of the Yang-Mills type combining weak and electromagnetic interactions, and this theory was proved to be renormalizable by G. 't Hooft in 1971. In 1974, the Glashow-Weinberg-Salam theory was combined with a new gauge theory of strong interactions, called quantum chromodynamics, and the synthesis became known as the standard model. *See* ELECTROWEAK INTERACTION; HIGGS BOSON; QUANTUM CHROMODYNAMICS; STANDARD MODEL; SYMMETRY BREAKING.

The electroweak sector of the standard model contains two gauge symmetry groups, $SU(2)_L$ and U(1), with coupling constants g and g', respectively. [The subscript L refers to "left-handed fermions" (and "right-handed antifermions"), meaning that parity violation is built into the theory from the beginning.] As mentioned previously, there is a triplet of gauge fields associated with SU(2) that have "SU(2) charges" $+g$, $-g$, and 0, while there is only a single neutral field in U(1). In the standard model, the physical electromagnetic and neutral weak fields are constructed as orthogonal linear combinations of the neutral SU(2) and U(1) fields. These linear combinations are characterized by a weak mixing angle, $\theta_W = \tan^{-1}(g'/g)$, which is found to be $\theta_W = 0.49$ rad. Various relationships [Eqs. (7)–(9)] between g, θ_W, the masses m_W and m_Z,

$$\sqrt{4\pi\alpha} = g \sin\theta_W \tag{7}$$

$$m_z = \frac{m_W}{\cos\theta_W} \tag{8}$$

$$G_F = \frac{g^2}{2^{5/2} m_W^2} \tag{9}$$

Fermi's constant G_F, and the fine-structure constant

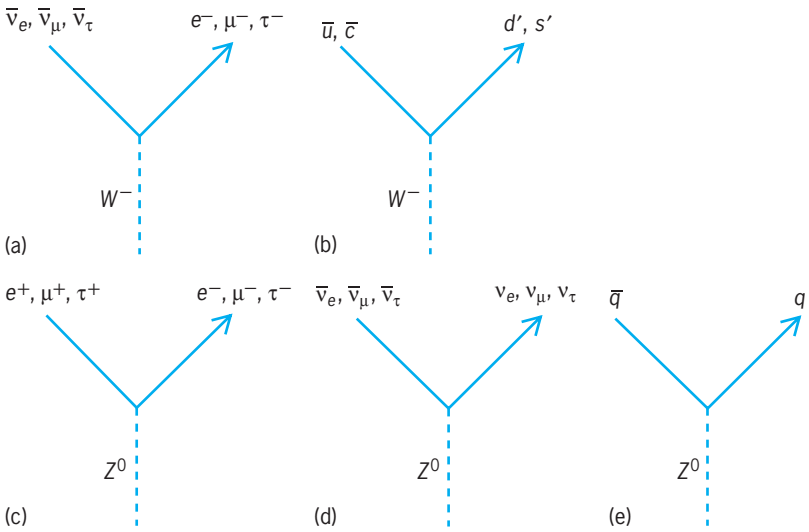


Fig. 3. Decay modes of intermediate vector bosons W^- and Z^0. (*a*) W^- decay to a charged lepton and an antineutrino. (*b*) W^- decay to a quark and an antiquark. Here, d' and s' refer to the weak eigenstates obtained from the quark mass eigenstates *d*, *s*, and *b* by the CKM matrix transformation. (The decay $W^- \rightarrow \overline{t}b'$ cannot occur because the mass of the top quark is too large. Instead the chief mode of decay of a top quark is $t \rightarrow W^+b$.) (*c*) Z^0 decay to a charged lepton–antilepton pair. (*d*) Z^0 decay to a neutrino–antineutrino pair. (*e*) Z^0 decay to a quark–antiquark pair. Here, q refers to a *u*, *d*, *c*, *s*, or *b* quark, while $\overline{q}$ refers to the corresponding antiquark. (The Z^0 cannot decay to a $t\overline{t}$ pair, again because the mass of the top quark is too large.)

$\alpha = 1/137.036$ (which is the fundamental parameter of quantum electrodynamics) follow from the mixing and from the requirement that the new theory conform to the known and valid results of quantum electrodynamics and the V-A law for low-energy charged weak interactions. The standard model also predicts the existence and detailed properties of neutral weak interactions, which were first observed in 1973, about 5 years after the theory's creation.

Verification of the standard model. Experiments carried out between 1973 and 1980 confirmed the correctness of the theory for low-energy neutral weak interactions (that is, in reactions where the energy transmitted by the intermediate boson Z^0 is small compared to its rest energy). These included neutrino–electron scattering, neutrino–nucleon scattering, parity violation in atoms, and scattering of polarized electrons on nucleons. The latter two cases involve interference between the weak and electromagnetic interactions. Then, in 1982–1983 at the European Center for Nuclear Research (CERN), the intermediate bosons $W^\pm$ and Z^0 were directly produced in energetic proton–antiproton ($p\overline{p}$) collisions, and detected by means of their decay products (**Fig. 3**). Subsequently, experiments were carried out at the Large Electron-Positron Collider (LEP) at CERN between 1989 and 2000, and at the Stanford Linear Collider (SLC) of the Stanford Linear Accelerator Center (SLAC), at more or less the same time, to investigate the properties of Z and W bosons in great detail. Precise measurements of the W mass were also undertaken at the Tevatron at Fermilab. In these experiments, the predictions of the standard electroweak model concerning the masses of W and Z, their couplings to leptons and quarks, and their relation to the fine-structure constant, including quantum-field-theoretic "radiative" corrections to first nonvanishing ("one-loop") order, were verified to remarkable precision. These radiative corrections include effects due to the top quark and the Higgs boson. They may thus be employed in conjunction with experimental data to determine the top quark mass m_t with small uncertainty, and to put useful lower and upper bounds on the Higgs boson mass m_H. The value of m_t obtained in this way agrees with the value from direct observations of the top quark that have been made since 1995–1996 at Fermilab. Incontrovertible observations of the Higgs boson do not yet exist, but a lower limit exists on m_H from the negative results of experimental searches so far. This limit is consistent with the limits obtained from radiative corrections (**Table 3**). It is widely expected that direct detection of the Higgs boson will be achieved by experiments at the Large Hadron Collider (LHC) at CERN, which should begin operation in 2007. *See* INTERMEDIATE VECTOR BOSON; PARTICLE ACCELERATOR.

Cabibbo-Kobayashi-Maskawa matrix. Although the fundamental quark states for strong interactions are as shown in Table 1, extensive data on charged weak decays of mesons and baryons from diverse experiments reveal the necessity of a unitary transformation of quark states *d*, *s*, and *b* to a new basis suitable for charged weak interactions. The first evidence for this was uncovered in the 1960s by N. Cabibbo, at a time when only the three "light" quarks, *u*, *d*, and *s* (and their corresponding antiquarks), were recognized. Cabibbo described the necessary linear transformation of *d* and *s* quark states by a single

TABLE 3. Some important physical quantities in weak interactions

Lepton parameters

Electron mass	$m_e = 0.51099892 \pm 0.00000004$ MeV/c^2
Muon mass	$m_\mu = 105.658369 \pm 0.000009$ MeV/c^2
Muon mean life	$\tau_\mu = (2.19703 \pm 0.00004) \times 10^{-6}$ s
Tau lepton mass	$m_\tau = 1776.99^{+0.29}_{-0.26}$ MeV/c^2
Tau mean life	$\tau_\tau = (2.906 \pm .011) \times 10^{-13}$ s

Direct mass limits on neutrinos from conservation of energy and momentum in various decays and reactions

$m(\nu_e) \leq 3$ eV/c^2
$m(\nu_\mu) \leq 0.19$ MeV/c^2
$m(\nu_\tau) \leq 18.2$ MeV/c^2

Number of neutrino families from Z^0 decay width (LEP data): 2.994 ± 0.012

Quark masses

$m_u = 1.5 - 4.5$ MeV/c^2
$m_d = 5 - 8.5$ MeV/c^2
$m_s = 80 - 155$ MeV/c^2
$m_c = 1.290^{+0.040}_{-0.045}$ GeV/c^2
$m_b = 4.206 \pm 0.031$ GeV/c^2
$m_t = 178.0 \pm 4.3$ GeV/c^2 (direct observation at Fermilab)
$= 177.9 \pm 4.4$ GeV/c^2 (from standard model radiative corrections)

Intermediate boson masses

$m_\gamma < 6 \times 10^{-17}$ eV/c^2
$m_W = 80.425 \pm 0.038$ GeV/c^2
$m_Z = 91.1876 \pm 0.0021$ GeV/c^2

Higgs boson mass

$m_H \geq 114.4$ GeV/c^2 (from negative results of direct searches for the Higgs boson)
$m_H = 113^{+56}_{-40}$ GeV/c^2 (from standard model radiative corrections)

"rotation angle." In 1973 M. Kobayashi and T. Maskawa proposed a generalization of these ideas to three generations of quarks (before direct experimental evidence for the third generation was available).

Assuming that there are three and only three quark generations, the transformation of d, s, and b states to the new basis, d', s', and b', suitable for charged weak interactions, is implemented by means of a 3×3 unitary matrix, called the Cabibbo-Kobayashi-Maskawa (CKM) matrix, as in Eq. (10). It may be

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \tag{10}$$

shown quite generally that the various matrix elements V_{ij} can be expressed in terms of three rotation angles, θ_{12}, θ_{23}, and θ_{13} (which generalize the original Cabibbo angle), and one complex phase δ. In the standard model, it is assumed that this phase alone accounts for the phenomenon of *CP* violation. No reliable theoretical model exists to calculate the matrix elements V_{ij} from first principles, and their origin is still a mystery.

CPT invariance and CP violation. All experimental evidence is consistent with *CPT* (combined charge conjugation, parity, and time reversal) invariance of strong, electromagnetic, and weak interactions, and there are compelling theoretical reasons to believe that *CPT* invariance is a universally valid symmetry. After *P* and *C* violation of the weak interaction was discovered in 1956–1957, it appeared for some years that the combined symmetry *CP* was also valid. However, it was found in 1964 that *CP* invariance is violated in certain weak decays of neutral K mesons. *See* CPT THEOREM.

The neutral K mesons, K^0 and $\overline{K}^0$, are charge conjugates of one another (their valence quark compositions are $K^0 = d\overline{s}$, $\overline{K}^0 = \overline{d}s$) and they possess definite strangeness eigenvalues ($+1$ and -1, respectively). Hence, it is appropriate to think in terms of the states K^0 and $\overline{K}^0$ when considering the strong and electromagnetic interactions, which conserve strangeness. However, K^0 and $\overline{K}^0$ do not possess definite lifetimes for weak decay, nor do they possess definite masses, because the charged weak interactions do not conserve strangeness. Instead, there exist two independent linear combinations of K^0 and $\overline{K}^0$ states, called $K_S^0$and K_L^0, which do have definite (and distinct) lifetimes and masses. This causes the phenomenon of K^0–$\overline{K}^0$ mixing or strangeness oscillations, first described in the 1950s by Gell-Mann and A. Pais, and independently by T. Nakano and K. Nishijima. The short-lived K_S^0 decays in only two significant modes, $\pi^+\pi^-$and $\pi^0\pi^0$, and each of these final states has *CP* eigenvalue $+1$. There are many known modes of decay of K_L^0, including the fully allowed decay to $\pi^+\pi^-\pi^0$, a final state which predominantly has *CP* eigenvalue -1. Thus, if *CP* were conserved, the decays $K_L^0 \to \pi^+\pi^-$ and $K_L^0 \to \pi^0\pi^0$ would be strictly forbidden. However, J. Cronin, V. Fitch, and coworkers observed such decays with small but finite probability in 1964, which implies *CP* violation, and these results have been verified in great detail in the years since.

The data reveal that the largest *CP*-violating effect occurs in the decay amplitudes of K_L^0 to various final states ("indirect" *CP* violation) but there is also a small so-called direct *CP* violation in K^0–$\overline{K}^0$ mixing. The standard model allows for the possibility of direct as well as indirect *CP* violation in kaon decay, but theoretical prediction of direct *CP* violation within the standard model framework has large uncertainties due to the strong interaction. Hence it is not yet possible to conclude that the standard model satisfactorily accounts for *CP* violation in neutral kaon decay.

CP violation has also been observed in the $B^0 - \overline{B}^0$ system (at the BABAR installation at SLAC, and at the BELLE installation in Japan). In these experiments, e^+e^- collisions produce $b\overline{b}$ quark pairs which "hadronize" to form a $B^0 = b\overline{d}$, $\overline{B}^0 = \overline{b}d$ meson pair (or $B_s^0 = b\overline{s}$ and $\overline{B}_s^0 = \overline{b}s$ mesons) with equal and opposite momenta in the center-of-momentum frame. Because of "particle mixing" through the weak interaction, which is somewhat analogous to that in the neutral kaon system, each meson evolves continuously into its charge conjugate and back again as time elapses. However, at some point one of the mesons decays. Observation of this "tagged" decay serves to identify whether the tagged meson was B^0 or $\overline{B}^0$ when it decayed, implying that the surviving meson is of the other type. The "tagging" also starts a clock for measuring the time-dependent probability for the surviving meson to decay to a final state with definite *CP* eigenvalue (for example, $J/\psi K_L^0$). In this way a *CP*-violating asymmetry in the time-dependent decay probabilities can be measured. The resulting effects are large, and for the decay modes observed so far, they are in agreement with predictions of the standard model. However, there is reason to believe that some rare decay modes of B^0 or B_s^0 mesons, yet to be observed with precision in these experiments, may exhibit *CP*-violating effects that must be accounted for by descriptions that go beyond the standard model.

Given *CPT* invariance, *CP* violation necessarily implies *T* violation. Thus efforts have been made for many years to observe *T* violation elsewhere than in the neutral kaon or neutral bottom meson systems. One particular case is of special interest: the search for permanent spin electric dipole moments of the neutron and the electron.

These particles, among others, possess spin $^1/_2$ and have spin magnetic dipole moments. However, they cannot possess electric dipole moments unless both parity (*P*) and time reversal (*T*) invariance are violated. Of course, parity is indeed violated in the weak interaction and *CP* violation (equivalent to *T* violation) also occurs. Thus, electric dipole moments can exist through radiative corrections to the electromagnetic interaction arising from the weak interaction and the *CP*-violating mechanism. However, it can be shown that so long as *CP* violation is

described by the standard model, the electric dipole moments of the neutron and the electron would be far too small by many orders of magnitude for any practical experiment to detect. On the other hand, these electric dipole moments could be within experimental reach according to various plausible models of new physics beyond the standard model (for example, supersymmetric models of various types). Motivated by this possibility, sensitive searches for the neutron and electron electric dipole moments have been carried out, and new searches are being prepared. However, only upper limits on electric dipole moments have been obtained so far. *See* DIPOLE MOMENT; TIME REVERSAL INVARIANCE; SUPERSYMMETRY.

Weak interactions of neutrinos. According to the standard model, neutrinos and antineutrinos have zero mass, and ν_e, ν_μ, ν_τ have helicity $h = -1$ (are left-handed), whereas $\overline{\nu}_e$, $\overline{\nu}_\mu$, $\overline{\nu}_\tau$ have $h = +1$ (are right-handed). Is this description correct? Are neutrinos really massless, like photons, or do they possess small but nonzero mass? If at least one neutrino species has mass, the possibility exists that the neutrino weak interaction eigenstates ν_e, ν_μ, ν_τ are not the same as the neutrino states of definite mass (the latter are called ν_1, ν_2, ν_3). This idea of neutrino mixing is suggested by analogy with quark mixing, where, as noted earlier, the quark weak eigenstates, d', s', and b' are not the same as the mass eigenstates, d, s, and b, but are related to the latter by the CKM matrix of Eq. (10).

In fact, in four distinct experiments performed since 1998 with atmospheric neutrinos, accelerator neutrinos, and reactor neutrinos in Japan, and with solar neutrinos in Canada, compelling evidence has been found that at least two neutrino species do have nonzero mass, and that neutrino mixing actually occurs. The mixing is described by a unitary 3×3 matrix U (the "lepton mixing matrix") in Eq. (11).

$$\begin{pmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{pmatrix} = \begin{pmatrix} U_{e1} & U_{e2} & U_{e3} \\ U_{\mu 1} & U_{\mu 2} & U_{\mu 3} \\ U_{\tau 1} & U_{\tau 2} & U_{\tau 3} \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix} \qquad (11)$$

The matrix U, like the CKM matrix, is characterized by three "lepton" mixing angles, θ_{12}, θ_{23}, and θ_{13}, and a phase δ, although the numerical values of the leptonic quantities are not the same as the analogous CKM quantities.

Neutrino mixing causes the remarkable phenomenon of neutrino oscillations, in which a neutrino created in a definite weak eigenstate at some initial time $t = 0$ transforms into a superposition of all neutrino weak eigenstates as time elapses. For example, a solar neutrino is born in the core of the Sun as ν_e, but by the time it arrives at a detector on Earth it is in a superposition of states ν_e, ν_μ, and ν_τ. The probability to find such a neutrino at the detector in the state ν_e is therefore less than unity, and a detector that is sensitive only to ν_e will thus register fewer counts than it would have if neutrino oscillations did not occur. On the other hand, a detector that is equally sensitive to each of the neutrino weak eigenstates, ν_e, ν_μ, and ν_τ, will not suffer any such loss of counts. Experiments carried out at the Sudbury Neutrino Observatory (SNO) in Canada using both schemes of detection thus demonstrated the existence of solar neutrino oscillations and at the same time verified that the solar neutrino flux is as predicted by detailed theoretical models of the Sun.

The leptons (e^-, μ^-, τ^-) and their respective antileptons (e^+, μ^+, τ^+) have opposite electric charges and opposite magnetic moments, and are therefore different from one another. Because of this difference they are called Dirac particles. Each neutrino of a given type may also be different from its corresponding antineutrino (may be a Dirac neutrino). However, neutrinos are electrically neutral, and it is therefore theoretically possible for a neutrino of a given type to be identical to its corresponding antineutrino (in which case it is referred to as a Majorana neutrino). The distinction between these two possibilities has physical meaning only if the neutrino has nonzero mass. Thus, now that the existence of neutrinos with mass has been demonstrated, the question "Are neutrinos of the Dirac type or the Majorana type?" becomes especially significant. It can be addressed experimentally as follows. Nuclei exist for which ordinary beta decay is energetically forbidden or highly suppressed by conservation of angular momentum, but which are unstable for double beta decay, a second-order process in which two electrons are emitted. Ordinarily these two electrons are accompanied by two antineutrinos ("$\beta\beta_2$" decay). However, neutrinoless double beta decay ("$\beta\beta_0$" decay), although very difficult to observe, is also possible, but only if neutrinos are of the Majorana type. Experimental searches for $\beta\beta_0$ decay are being carried out, but they have not yet yielded definitive results.

Eugene D. Commins

Bibliography. T.-P. Cheng and L.-F. Li, *Gauge Theory of Elementary Particle Physics*, Oxford University Press, 1984, reprinted with corrections, 1988; E. D. Commins and P. H. Bucksbaum, *Weak Interactions of Leptons and Quarks*, Cambridge University Press, 1983; M. Fukugita and T. Yanagida, *Physics of Neutrinos*, Springer, Berlin, 2003; F. Halzen and A. D. Martin, *Quarks and Leptons: An Introductory Course in Modern Particle Physics*, Wiley, New York, 1984; B. R. Holstein, *Weak Interactions in Nuclei*, Princeton University Press, 1989; P. Langacker (ed.), *Precision Tests of the Standard Electroweak Model*, World Scientific, Singapore, 1995; L. B. Okun, *Particle Physics: The Quest for the Substance of Substance*, Harwood, London, 1985.

Weakly interacting massive particle (WIMP)

A hypothetical elementary particle that might make up most of the matter in the universe, and that is also predicted to exist in supersymmetry theory. Most matter is detected only through its gravitational effects; this "dark matter" has not been observed to emit, absorb, or reflect light of any wavelength. The total amount of dark matter appears to be approximately ten times as great as all the ordinary matter in the universe, and about one hundred times as great

as all the visible matter. The nature of the dark matter is not yet known, although many experiments are under way to try to discover it directly or indirectly. *See* COSMOLOGY; UNIVERSE.

Almost all the currently available data in elementary particle physics can be accounted for by a theory called the standard model, in which matter is made of quarks (the building blocks of protons and neutrons) and leptons (including electrons and neutrinos), while the strong, weak, and electromagnetic forces are transmitted by particles like the photon (the carrier of electromagnetic forces). However, the standard model does not predict the existence of any particle—say, χ—that could be the dark matter. Most efforts to go beyond the standard model of particle physics have been based on the idea of supersymmetry, and most versions of supersymmetry predict that there will be a stable weakly interacting massive particle (WIMP) that would be a natural candidate for the χ particles. Dark matter made of WIMPs would be "cold" dark matter (CDM), and a version of CDM theory has become the standard theory of structure formation in cosmology. Its predictions agree very well with the observed properties of the universe on large scales, but there may be disagreements on scales smaller than galaxies. *See* ELEMENTARY PARTICLE; STANDARD MODEL; SUPERSYMMETRY.

Evidence for weakly interacting dark matter. There is now abundant evidence for dark matter around galaxies and clusters of galaxies, and on larger scales in the universe. In the solar system, the more distant a planet is from the Sun, the slower its velocity around the Sun. This is as expected from standard gravity theory and mechanics, since almost all the gravitating mass in the solar system is in the Sun at its center. But gas and satellites at large distances from galaxies have orbital velocities similar to those at smaller distances from the center, which indicates that most of the mass in the galaxy must not be near the center, where most stars are, but in a roughly spherical dark matter halo that extends to perhaps ten times the optical size of the galaxy and has a mass at least ten times that of all the stars. Confirmation of the existence of such dark-matter halos has come from gravitational lensing observations, showing that light from more distant galaxies is bent by the gravity of nearer galaxies. *See* CELESTIAL MECHANICS; GALAXY, EXTERNAL; GRAVITATIONAL LENS; MILKY WAY GALAXY.

Large, nearly spherical halos are expected to form by gravitational collapse in the expanding universe. Once such a halo forms, the gas in it (mostly hydrogen and helium) can continue to collide and convert kinetic energy to radiation, sinking closer to the center of attraction. If the gas retains its angular momentum, it will spin faster as its radius decreases, and will ultimately form a disk. When much of the gas has formed into stars, such a galaxy would be a spiral, the most abundant type of large galaxy. Collisions between such disk galaxies may produce most elliptical galaxies and bulges of spiral galaxies. But if the dark matter can interact only weakly—that is, via the weak interactions, carried by the gauge particles (generalizations of the photon) W^+, W^-, and Z^0—but not strongly or electromagnetically, that would explain why χ particles cannot lose energy by radiation, but must form extended dark halos. Weakly interacting χ particles would also not be concentrated in minerals on Earth, or anywhere else in the disk of the Milky Way Galaxy for that matter, in agreement with observations. *See* INTERMEDIATE VECTOR BOSON; WEAK NUCLEAR INTERACTIONS.

There is also much evidence for dark matter in clusters of galaxies. The astronomer Fritz Zwicky pointed out in 1933 that the galaxies in one nearby cluster were moving at such high speeds that they would not be held together gravitationally unless there was much more mass than was indicated by the light from their stars. This same was subsequently found to be true of other clusters. Later, similar conclusions were reached from x-ray observations and gravitational lensing observations of clusters. *See* X-RAY ASTRONOMY.

Measurements of the typical angular sizes of structures in the cosmic background radiation—the heat radiation of the big bang—provide information about the composition of the universe, and again lead to the conclusion that most of the matter in the universe is dark matter. All of these data indicate that the total amount of dark matter is about 30–40% of the critical density required to close the universe, while the total amount of ordinary matter is only about 4–5% of the critical density, and the amount of visible matter is less than half a percent of the critical density. The density in units of critical density is usually written as Ω, a dimensionless number; the universe will expand forever if the total average density $\Omega < 1$, and will eventually collapse if $\Omega > 1$. Expressed this way, the average, or cosmological, dark-matter density is $\Omega_{dm} \approx 0.3$–0.4, the density of all ordinary matter (baryons) is $\Omega_b \approx 0.04$–0.05, while the density of visible matter is only $\Omega_{vis} \approx 0.005$. *See* BIG BANG THEORY; COSMIC BACKGROUND RADIATION.

Nature of dark matter. Since about 1980, when the evidence for dark matter became convincing, there have been various theories proposed concerning its possible composition. That dark matter is invisible suggests that the χ particles are electrically neutral. The leading candidates for the dark-matter particle are axions and supersymmetric WIMPs, since both were predicted on the basis of particle physics theories that were motivated by laboratory experiments, and it was only subsequently appreciated that they could solve the dark-matter mystery.

Supersymmetric WIMPs. Supersymmetry is the hypothesis that there is a relationship between the two known classes of particles, bosons and fermions. The bosons, such as the photon and the graviton, are responsible for the forces of nature. The fermions—including the atomic building blocks, the proton, neutron, and electron—are the matter particles. According to supersymmetry, for every kind of boson in the universe, there must also be a corresponding fermion with the same electric charge and very similar interactions with other particles. For example, the supersymmetric analog of the photon has been named the photino, and like the photon, it would be

electrically neutral. Since these hypothetical supersymmetric partner particles have not been discovered yet, if supersymmetry is right their masses must be too large for them to have been produced at current particle accelerators. Thus far, the evidence for supersymmetry is only indirect, but if the theory is right many supersymmetric partner particles should be produced at accelerators such as the Large Hadron Collider (LHC) being built in Geneva, Switzerland and scheduled to start operating in 2006. *See* PARTICLE ACCELERATOR; PHOTON; QUANTUM STATISTICS.

In most versions of supersymmetry, there is a new conserved quantum number called R-parity, and the ordinary particles all have R-parity equal to $+1$, while the superpartner particles all have R-parity equal to -1. Thus while heavier superpartner particles can decay into lighter ones (plus ordinary particles), the lightest superpartner (LSP) must be stable, since there is no lighter negative R-parity particle into which it can decay. Thus, the universe should be full of LSP particles if the temperatures in the early universe was high enough to produce supersymmetric partner particles—which is the case according to most theories of the early universe. *See* CONSERVATION LAWS (PHYSICS); INFLATIONARY UNIVERSE COSMOLOGY.

This argument was first published in 1982, with the gravitino (the superpartner of the graviton) as the LSP. In currently favored versions of supersymmetry theory, however, the LSP is a neutralino (the superpartner of some other neutral particle) such as the photino; the theory of such dark-matter particles was published in 1984. Neutralinos would be massive but have interactions like those of neutrinos and other particles that interact only via the weak interaction (and gravity); thus they would be supersymmetric WIMPs. These hypothetical particles would mostly go through whole planets without ever interacting. Nevertheless, it is feasible to search for such WIMPs with very sensitive new instruments. *See* GRAVITON.

Searching for WIMPs. Efforts to detect WIMPs directly are based on detecting their scattering from nuclei. The weak nature of the interactions of WIMPs means that the probability of their interacting is very small. But for the few that do interact, the amount of energy delivered in the collision is large enough to detect with sensitive instruments. Since the WIMPs are supposed to be gravitationally bound to the Milky Way Galaxy, their velocities v are approximately 300 km/s (180 mi/s), a thousandth the speed of light, c. It follows that the ratio of their kinetic energy ($= mv^2/2$ for these nonrelativistic particles of mass m) to their rest mass energy (mc^2) will be ${}^1/_2(v/c)^2 \approx 0.5 \times 10^{-6}$. If their rest mass is approximately one hundred times the mass of a proton—the sort of mass expected since the charged superpartners have not yet been produced at particle accelerators—then their kinetic energy will be about 50 kiloelectronvolts. All this kinetic energy would be transferred as recoil energy to an equal-mass nucleus struck head on. This would have two different sorts of effects in a crystal: vibrations and electronic excitations. The vibrations would quickly become random heat vibrations, and these can be detected in crystals cooled to a small fraction of a degree above absolute zero. The excited electrons could be detected as a current if the crystal is a semiconductor, or as emitted light if the crystal is transparent and has the appropriate electronic structure. It is very useful to detect both vibrations and electron excitations, since the main background of events that could be mistaken for WIMP interactions is scattering of electrons in the detector by energetic photons (Compton scattering). Since the background events would put more energy into electron excitations than into crystal lattice vibrations, this can be used to help tell whether a possible signal is coming from WIMP scattering. Many experiments of this type are being built around the world. It is desirable to put these detectors deep underground to shield them from cosmic rays. *See* COMPTON EFFECT; COSMIC RAYS; LATTICE VIBRATIONS; PARTICLE DETECTOR; SCINTILLATION COUNTER.

WIMPs can also be detected indirectly, for example by looking for particles coming from their annihilation. Supersymmetric WIMPs (neutralinos) are their own antiparticles, and their annihilation products include energetic photons (gamma rays) and neutrinos, as well as particle-antiparticle pairs, such as electrons and positrons, or protons and antiprotons. One possible signal is gamma rays from WIMP annihilation at the center of the Milky Way Galaxy, where the WIMP density is elevated because of the structure of dark-matter halos, and perhaps further elevated (or perhaps alternatively decreased—this is a controversial issue) by the presence of a black hole at the galactic center. The Gamma-ray Large Area Space Telescope (GLAST), scheduled for launch in 2006, will look for this signal with greatly increased sensitivity. Another important signal that has been searched for is neutrinos coming from WIMP annihilation in the center of the Sun or the center of the Earth, where some will settle after being slowed down by scattering off atomic nuclei in the Sun or Earth. *See* ANTIMATTER; BLACK HOLE; GAMMA-RAY ASTRONOMY; NEUTRINO ASTRONOMY.

WIMPs are also expected to be produced at accelerators such as the LHC from rapid decays of heavier supersymmetric partner particles, and this could be where they are discovered first if they are not seen before that in direct or indirect search experiments. Failure to see supersymmetric particles at LHC energies would mean that current ideas about supersymmetry are wrong.

WIMPs as cold dark matter. The CDM theory was developed to describe how galaxies and clusters of galaxies would form in a universe in which the dark matter is WIMPs or other particles that would have been "cold" in the early universe—that is, moving so sluggishly that their motion can be neglected. "Hot" dark matter (HDM) had earlier been considered. HDM refers to light particles like neutrinos, thought to have rest energies of perhaps a few electronvolts. Such particles would be moving at nearly the speed of light in the hot early universe, and this would inhibit formation of galaxies. The first

structures to form in an HDM universe would be objects with the mass of a supercluster of galaxies, and galaxies would have to form subsequently through their fragmentation. But superclusters are still forming today, while most galaxies are old. That HDM predicted the wrong sequence of cosmogony led to its downfall, while CDM is a hierarchical structure formation theory in which smaller things form earlier and merge to form larger and larger structures, up to clusters of galaxies, the largest gravitationally bound objects. *See* NEUTRINO.

The distribution and properties of galaxies predicted by the CDM theory agree beautifully with observations. Only at the centers of galaxies is there a possible discrepancy between theory and observation: theory predicts a density cusp at the center of dark-matter halos, while some observations suggest a more uniform dark-matter density there. Although such a cusp would involve only a tiny fraction of a galaxy's mass, this possible discrepancy is a serious concern for the theory that the dark matter is mostly WIMPs. Joel R. Primack

Bibliography. G. Blumenthal et al., Formation of galaxies and large-scale structure with cold dark matter, *Nature*, 311:517–525, 1984; G. Jungman, M. Kamionkowski, and K. Griest, Supersymmetric dark matter, *Phys. Rep.*, 267:195–373, 1996; B. Moore, Evidence against dissipationless dark matter from observations of galaxy halos, *Nature*, 370:629–231, 1994; J. R. Primack, D. Seckel, and B. Sadoulet, Detection of cosmic dark matter, *Annu. Rev. Nucl. Part. Sci.*, 38:751–807, 1988.

Wear

The removal of material from a solid surface as a result of sliding action. It constitutes the main reason why the artifacts of society (automobiles, washing machines, tape recorders, cameras, clothing) become useless and have to be replaced. There are a few uses of the wear phenomenon, but in the great majority of cases wear is a nuisance, and a tremendous expenditure of human and material resources is required to overcome the effects.

For reasons which are hard to explain, wear did not emerge as a technical field until the 1950s. Before then, the feeling seems to have been that wear was an inevitable accompaniment of sliding, and that nothing could be done about it. Perhaps the key event which, rather indirectly, led to people's change in attitude was the building of nuclear reactors as part of the atomic bomb program in the 1940s. These reactors produced radioisotopes of the common engineering metals (iron, copper, chromium, and so on), and these allowed the use of extremely sensitive radiotracer techniques for measuring wear. For example, it became possible to measure the very small amount of wear which occurs when one metal is pressed with a force of a kilogram against another and then slides over it.

Thus, for the first time it became possible to measure wear while it was occurring, rather than by before-and-after measurements. As result of this research, it was established by about 1950 that there were four principal forms of wear—adhesive, abrasive, corrosive, and surface fatigue—and it soon became possible to work out their mechanisms and to express the amount of wear in quantitative terms.

Adhesive wear. This is the only universal form of wear, and in many sliding systems it is also the most important. It arises from the fact that, during sliding, regions of adhesive bonding, called junctions, form between the sliding surfaces. If one of these junctions does not break along its original interface, then a chunk from one of the sliding surfaces will have been transferred to the other surface. In this way, an adhesive wear particle will have been formed. Initially adhering to the other surface, adhesive particles soon become loose and can disappear from the sliding system.

The volume of adhesive wear is governed by the equation $V = kLx/3p$, where V is the total volume of adhesive wear, k is a nondimension constant called the wear coefficient, x is the total distance of sliding, and p is the indentation hardness of the surface expressed as a stress. Typical values of k are given in **Table 1**.

A very great reduction in wear, by factors of up to a million, can be produced in metallic sliding systems by using a good lubricant. Also, there is a great advantage in making unlubricated sliding systems nonmetallic, and well-lubricated systems metallic.

Adhesive wear causes two types of failure, one being a wear-out mode which occurs after long periods of sliding because too much material has been removed, and a seizure mode which occurs in systems which generate wear particles larger than the clearance, thus producing jamming. Since many material combinations give wear particles larger than 10 micrometers, it is dangerous to reduce the clearance of any sliding system below this value. *See* FRICTION.

Abrasive wear. This is the wear produced by a hard, sharp surface sliding against a softer one and digging out a groove. The abrasive agent may be one of the surfaces (such as a file), or it may be a third component (such as sand particles in a bearing abrading material from each surface). Abrasive wear, like adhesive wear, obeys the equation given above; typical values of k are given in **Table 2**.

It will be seen that abrasive wear coefficients are large compared to adhesive ones. Thus, the introduction of abrasive particles into a sliding system can greatly increase the wear rate; automobiles, for example, have air and oil filters to catch abrasive particles before they can produce damage.

TABLE 1. Values of the wear coefficient for adhesive wear

Surface condition	Metal on metal	Nonmetal on metal, or nonmetal on nonmetal
Clean	10^{-2} to 10^{-4}	10^{-5}
Poor to fair lubricant	10^{-4} to 10^{-6}	3×10^{-6}
Good lubricant	10^{-6} to 10^{-8}	10^{-6}

TABLE 2. Values of the wear coefficient for abrasive wear

Process	*k* value
Sharp file	2×10^{-1}
Sandpapering	5×10^{-2}
Loose abrasive grains	5×10^{-3}
Polishing	5×10^{-4}

Corrosive wear. This form of wear arises when a sliding surface is in a corrosive environment, and the sliding action continuously removes the protective corrosion product, thus exposing fresh surface to further corrosive attack. No satisfactory quantitative expression of corrosive wear yet exists, but when analyzing it in terms of the above equation, *k* values are obtained ranging all the way from less than 10^{-5} for surfaces in a gently corrosive environment, to above 10^{-2} for surfaces under severe corrosive attack. *See* CORROSION.

Surface fatigue wear. This is the wear that occurs as result of the formation and growth of cracks. It is the main form of wear of rolling devices such as ball bearings, wheels on rails, and gears. During continued rolling, a crack forms at or just below the surface and gradually grows until a large particle is lifted right out of the surface.

Uses. Most manifestations of wear are highly objectionable, but the phenomenon does have a few uses. Thus, a number of systems for recording information (pencil and paper, chalk and blackboard) operate via a wear mechanism. Some methods of preparing solid surfaces (filling, sandpapering, sandblasting) also make use of wear. *See* ABRASIVE.

Wear is often used to study the operation of sliding systems. Thus, by looking at the wear of someone's shoes, it can be seen whether the wearer is walking properly. In analyzing accidents and failures, one of the main procedures is to look for signs of wear in the wrong place. Wear is specially useful in this case because it can be used in systems that are no longer operative.

Ernest Rabinowicz

Bibliography. B. Bhushan and B. Gupta, *Handbook of Tribology: Materials, Coatings, and Surface Treatments*, 1991; E. Rabinowicz, *Friction and Wear of Materials*, 2d ed., 1995; N. P. Suh, *Tribophysics*, 1986; N. P. Suh and N. Saka (eds.), *Fundamentals of Tribology*, 1980; R. B. Waterhouse (ed.), *Fretting Fatigue*, 1991.

Weasel

The common name for 14 species belonging to the family Mustelidae. These carnivores are native to every continent except Australia and Antarctica (see **table**). They are also absent from Madagascar and most oceanic islands.

Morphology. Weasels have a long slender body and short legs (see **illustration**). Most are brown or reddish-brown on the back and whitish or yellowish on the underside, although *Poecilictis* and *Poecilogale* are black with white spots and/or stripes. During winter in northern areas, the fur of some species, such as the long-tailed and short-tailed weasels, changes entirely to white except for the tip of the tail, which may remain black. The limbs possess five digits with nonretractile, curved claws. Ears are short and rounded. The dental formula is I 3/3, C 1/1, PM 3/3, M 1/2 × 2 for a total of 34 teeth. *See* DENTITION.

Long-tailed weasel (*Mustela frenata*). (*Alden M. Johnson* © *California Academy of Sciences*)

North American species. Three species occur in North America. The long-tailed weasel (*Mustela frenata*) has a head and body length ranging 203–260 mm (8–10 in.), a tail length of 76–152 mm (4–6 in.), and weighs 85–340 g (3–12 oz). The short-tailed weasel (*M. erminea*) has a head and body length between 170 and 325 mm (6.5 and 12.5 in.), a tail length of 42–120 mm (1.75–4.5 in.), and a weight between 42 and 365 g (1.5 and 12.5 oz). The least weasel (*M. nivalis*) is the smallest carnivore in North America. Adults are usually between 165 and 220 mm (6.5 and 8.5 in.) long including a tail

TABLE 1. Names and distribution of weasal species

Scientific and common name	Distribution
Mustela felipei (Colombian weasel)	South America
M. africana (tropical weasel)	South America
M. nivalis (least or common weasel)	North America, Europe, Asia, Africa
M. erminea (short-tailed weasel; ermine; stoat)	North America, Europe
M. frenata (long-tailed weasel)	North and South America
M. altacia (mountain weasel)	Asia
M. kathiah (yellow-bellied weasel)	Asia
M. sibirica (Siberian weasel)	Asia
M. lutreolina (Indonesian mountain weasel)	Sumatra, Java
M. nudipes (Malaysian weasel)	Malaysia, Borneo
M. strigidorsa (back-striped weasel)	Southern Asia
Lyncodon patagonicus (Patagonian weasel)	Argentina, Chile
Poecilictis libyca (North African striped weasel)	Africa
Poecilogale albinucha (African striped weasel)	Africa

25–45 mm (1–2 in.) long. They weigh between 25 and 65 g (1 and 2.3 oz). Male weasels are usually much larger than females.

Behavior and habitat. Weasels are very quick and their movements are difficult to follow. When running, the back is strongly arched at each bound and the tail is held straight out behind or at an angle. Weasels have keen senses of smell, sight, and hearing and are tireless hunters. They are voracious feeders, catching animals their own size or larger for food. They feed primarily on mice, rats, shrews, and moles, although earthworms, insects, frogs, lizards, rabbits, squirrels, snakes, and small birds may also be eaten. Weasels sever the prey's spinal column with one or two strong, swift bites at the base of the skull. The weasels' slim body and short legs enable it to hunt in the narrow openings of stone walls, under logs and rocks, and in rodent burrows. A weasel can follow a mouse to the end of the mouse's burrow and can squeeze through knotholes into chicken coops. Weasels raid farms and often kill more chickens than they need for food. As a result, many farmers dislike weasels even though they destroy farmyard pests. Although chiefly ground-dwelling animals, many weasels are excellent tree climbers and are able to raid birds' nests.

Weasels are more active at night than during the day, and they remain active in winter. They inhabit farmland as well as woodlands and swamps. Hedgerows and brushy field borders provide good habitat. Areas in the vicinity of water seem to be preferred. They make their dens in rock piles, under tree stumps, in sawmill slab piles, in abandoned rodent burrows, or occasionally in barns or other buildings. They sometimes store food in these dens and make nests from the fur, feathers, bones, and other remains of prey.

Reproduction. Most female weasels produce one annual litter of four to eight young. Ovulation is induced by copulation. Delayed implantation of the fertilized eggs occurs in many species. For example, embryos of the long-tailed weasel do not become implanted in the wall of the uterus for many months. Thus, whereas breeding normally occurs in July or August, the young are not born until April or early May of the following year. The embryos are implanted only 21 to 28 days before birth, even though the total gestation period may range from 205 to 337 days (average 279 days). Least weasels do not experience delayed implantation. They have a total gestation of approximately 35 days and may have up to 3 litters a year.

Weasels have numerous enemies including snakes, hawks, owls, foxes, cats, and humans. Like skunks, weasels possess well-developed anal scent glands that discharge a foul-smelling liquid called musk when threatened or attacked. Some long-tailed weasels have lived in the wild up to three years. Captives have lived at least five years. *See* CARNIVORA.

Donald W. Linzey

Bibliography. D. W. Linzey, *The Mammals of Virginia*, The McDonald & Woodward Publishing Company, 1998; D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford Limited, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., The Johns Hopkins University Press, 1999.

Weather

The state of the atmosphere, as determined by the simultaneous occurrence of several meteorological phenomena at a geographical locality or over broad areas of the Earth. When such a collection of weather elements is part of an interrelated physical structure of the atmosphere, it is termed a weather system, and includes phenomena at all elevations above the ground. More popularly, weather refers to a certain state of the atmosphere as it affects humans' activities on the Earth's surface. In this sense, it is often taken to include such related phenomena as waves at sea and floods on land.

An orderly association of weather elements accompanying a typical weather system of the Northern Hemisphere may be illustrated by a large anticyclone, or high-pressure region. In such a "high," extending over an area of hundreds of square miles, the usually gentle winds circulate clockwise around the high-pressure center. This system often brings fair weather locally, which implies a bright sunny day with few clouds. The temperature may vary widely depending on season and time of day. However, a cyclone or low-pressure region is frequently associated with a dark cloudy sky with driving rain (or snow) and strong winds which circulate counterclockwise about a low-pressure center of the Northern Hemisphere.

A weather element is any individual physical feature of the atmosphere. At a given locality, at least seven such elements may be observed at any one time. These are clouds, precipitation, temperature, humidity, wind, pressure, and visibility. Each principal element is divided into many subtypes. For a discussion of a characteristic local combination of several elements, as they might be observed at a U.S. Weather Bureau station, *See* WEATHER MAP.

The various forms of precipitation are included by international agreement among the hydrometeors, which comprise all the visible features in the atmosphere, besides clouds, that are due to water in its various forms. For convenience in processing weather data and information, this definition is made to include some phenomena not due to water, such as dust and smoke. Some of the more common hydrometeors include rain, snow, fog, hail, dew, and frost.

Both a physical (or genetic) and a descriptive classification of clouds and hydrometeors have been devised. The World Meteorological Organization, which among many other activities coordinates the taking of weather observations among the nations of the world, recognizes at least 36 cloud types and 100 classes of hydrometeors.

Certain optical and electrical phenomena have long been observed among weather elements, including lightning, aurora, solar or lunar corona, and

halo. *See* AIR MASS; ATMOSPHERE; ATMOSPHERIC GENERAL CIRCULATION; CLOUD; FRONT; METEOROLOGY; PRECIPITATION (METEOROLOGY); STORM; WEATHER OBSERVATIONS; WIND. Philip F. Clapp

Bibliography. E. Aguado, J. E. Burt, and J. Burt, *Understanding Weather and Climate*, 3d ed., 2003; C. D. Ahrens, *Meteorology Today With Infotrac: An Introduction to Weather, Climate, and the Environment*, 6th ed., 1999; R. C. Barry and R. J. Chorley, *Atmosphere, Weather, and Climate*, 8th ed., 2003; W. R. Cotton and R. A. Pielke, *Human Impacts on Weather and Climate*, 1995.

Weather forecasting and prediction

Processes for formulating and disseminating information about future weather conditions based upon the collection and analysis of meteorological observations. Weather forecasts may be classified according to the space and time scale of the predicted phenomena. Atmospheric fluctuations with a length of less than 100 m (330 ft) and a period of less than 100 s are considered to be turbulent. Prediction of turbulence extends only to establishing its statistical properties, insofar as these are determined by the thermal and dynamic stability of the air and by the aerodynamic roughness of the underlying surface. The study of atmospheric turbulence is called micrometeorology; it is of importance for understanding the diffusion of air pollutants and other aspects of the climate near the ground. Standard meteorological observations are made with sampling techniques that filter out the influence of turbulence. Common terminology distinguishes among three classes of phenomena with a scale that is larger than the turbulent microscale: the mesoscale, synoptic scale, and planetary scale.

The mesoscale includes all moist convection phenomena, ranging from individual cloud cells up to the convective cloud complexes associated with prefrontal squall lines, tropical storms, and the intertropical convergence zone. Also included among mesoscale phenomena are the sea breeze, mountain valley circulations, and the detailed structure of frontal inversions. Because most mesoscale phenomena have time periods less than 12 h, they are little influenced by the rotation of the Earth. The prediction of mesoscale phenomena is an area of active research. Most forecasting methods depend upon empirical rules or the short-range extrapolation of current observations, particularly those provided by radar and geostationary satellites. Forecasts are usually couched in probabilistic terms to reflect the sporadic character of the phenomena. Since many mesoscale phenomena pose serious threats to life and property, it is the practice to issue advisories of potential occurrence significantly in advance. These "watch" advisories encourage the public to attain a degree of readiness appropriate to the potential hazard. Once the phenomenon is considered to be imminent, the advisory is changed to a "warning," with the expectation that the public will take immediate action to prevent the loss of life.

The next-largest scale of weather events is called the synoptic scale, because the network of meteorological stations making simultaneous, or synoptic, observations serves to define the phenomena. The migratory storm systems of the extratropics are synoptic-scale events, as are the undulating wind currents of the upper-air circulation which accompany the storms. The storms are associated with barometric minima, variously called lows, depressions, or cyclones. The sense of the wind rotation about the storm is counterclockwise in the Northern Hemisphere, but clockwise in the Southern Hemisphere. This effect, called geostrophy, is due to the rotation of the Earth and the relatively long period, 3–7 days, of the storm life cycle. Significant progress has been made in the numerical prediction of synoptic-scale phenomena.

Planetary-scale phenomena are persistent, quasi-stationary perturbations of the global circulation of the air with horizontal dimensions comparable to the radius of the Earth. These dominant features of the general circulation appear to be correlated with the major orographic features of the globe and with the latent and sensible heat sources provided by the oceans. They tend to control the paths followed by the synoptic-scale storms, and to draw upon the synoptic transients for an additional source of heat and momentum. Long-range weather forecasts must account for the slow evolution of the planetary-scale circulations. To the extent that the planetary-scale centers of action can be correctly predicted, the path and frequency of migratory storm systems can be estimated. The problem of long-range forecasting blends into the question of climate variation. *See* ATMOSPHERE; METEOROLOGICAL INSTRUMENTATION; MICROMETEOROLOGY; WEATHER OBSERVATIONS. Joseph P. Gerrity

The synoptic method of forecasting consists of the simultaneous collection of weather observations, and the plotting and analysis of these data on geographical maps. An experienced analyst, having studied several of these maps in chronological succession, can follow the movement and intensification of weather systems and forecast their positions. This forecasting technique requires the regular and frequent use of large networks of data. *See* WEATHER MAP.

The synoptic forecasts are limited because they are almost exclusively based upon surface observations. The stratosphere, a layer of air generally located above levels from 5.5 to 8 mi (9 to 13 km) and characterized by temperatures remaining constant or increasing with elevation, acts as a cap on much of the world's weather. A trackable instrument suspended beneath a rubber balloon (the radiosonde) is routinely deployed from a network of stations. The tracking of the balloon allows for wind measurements to be taken at various elevations throughout its ascent. Such a network of more than 2000 upper-air stations around the world reports its observations twice daily. Satellite measurements and aircraft

observations supplement the radiosonde network and are utilized to improve the analysis and tracking of weather systems.

The basic laws of hydrodynamics and thermodynamics can be used to define the current state of the atmosphere, and to compute its future state. This concept marked the birth of the dynamic method of forecasting, though its practical application was realized only after 1950, when sufficiently powerful computers became available to produce the necessary calculations efficiently. The first conceptual model of the life cycle of a surface cyclone showed that cyclonic storms typically formed along fronts, that is, boundaries between cold and warm air masses. The model showed that precipitation is associated with active cold and warm fronts, and that this precipitation wraps around the cyclone as it intensifies. Though subsequent work demonstrated that fronts are not crucial to cyclogenesis, the Norwegian frontal cyclone model is still used in weather map analyses. *See* CYCLONE; FRONT.

It was not until the late 1930s that routine aerological observations had a substantial impact on forecasting. By the 1950s, qualitatively reasonable numerical forecasts of the atmosphere were produced. Numerical weather prediction techniques, in addition to being applied to short-range weather prediction, are used in such research studies as air-pollutant transport and the effects of greenhouse gases on global climate change. *See* AIR POLLUTION; GREENHOUSE EFFECT; JET STREAM; UPPER-ATMOSPHERE DYNAMICS.

Though numerical forecasts continue to improve, statistical forecast techniques, once used exclusively with observational data available at the time of the forecast, are now used in conjunction with numerical output to predict the weather. Statistical methods, based upon a historical comparison of actual weather conditions with large samples of output from the same numerical model, routinely play a role in the prediction of surface temperatures and precipitation probabilities.

The recognition that small, barely detectable differences in the initial analysis of a forecast model often lead to very large errors in a 12–48-h forecast has led to experiments with ensemble forecasting. This method uses results from several numerical forecasts to produce the statistical mean and standard deviation of the forecasts. Success with ensemble forecasting suggests it can be a useful tool in enhancing prediction skill and in assessing the atmosphere's predictability. *See* DYNAMIC METEOROLOGY.

The best weather forecasts result from application of the synoptic method to the latest numerical and statistical information. The forecaster has an ever-increasing number of valuable tools with which to work. Numerical forecast models, owing to faster computers, are capable of explicitly resolving mesoscale weather systems. Geostationary satellites allow for continuous tracking of such dangerous weather systems as hurricanes. The increased routine use of Doppler radar, automated commercial aircraft observations, and use of data derived from wind and temperature profiler soundings promise to give added capability in tracking and forecasting mesoscale weather disturbances. Very high-frequency and ultrahigh-frequency Doppler radars may be used to provide detailed wind soundings. These wind profilers, if located sufficiently close to one another, will allow for the hourly tracking of mesoscale disturbances aloft. Ground- and satellite-based microwave radiometric measurements are being used to construct temperature and moisture soundings of the atmosphere. The assimilation of such data at varying times in the numerical model forecast cycle offers the promise of improved prediction and improved utilization of data and products by forecasters assisted by increasingly powerful interactive computer systems. Medium-range forecasts, ranging up to 2 weeks, may be improved from knowledge of forecast skill in relationship to the form of the planetary-scale atmospheric circulation. *See* DOPPLER RADAR; MESOMETEOROLOGY; METEOROLOGICAL RADAR; METEOROLOGICAL ROCKET; METEOROLOGICAL SATELLITES; RADAR METEOROLOGY; SATELLITE METEOROLOGY.

John R. Gyakum

Numerical Weather Prediction

Numerical weather prediction is the prediction of weather phenomena by the numerical solution of the equations governing the motion and changes of condition of the atmosphere.

The laws of motion of the atmosphere may be expressed as a set of partial differential equations relating the temporal rates of change of the meteorological variables to their instantaneous distribution in space. These equations are developed in dynamic meteorology. In principle, a prediction for a finite time interval can be obtained by summing a succession of infinitesimal time changes of the meteorological variables, each of which is determined by their distribution at a given instant of time. However, the nonlinearity of the equations and the complexity and multiplicity of the data make this process impossible in practice. Instead, it is necessary to resort to numerical approximation techniques in which successive changes in the variables are calculated for small but finite time intervals over a domain spanning part or all of the atmosphere. Even so, the amount of computation is vast, and numerical weather prediction remained only a dream until the advent of the modern computer.

The accuracy of numerical weather prediction depends on (1) an understanding of the physical laws of atmospheric behavior; (2) the ability to define through observations and analysis the state of the atmosphere at the initial time of the forecast; and (3) the accuracy with which the solutions of the continuous equations describing the rate of change of atmospheric variables are approximated by numerical means. The greatest success has been achieved in predicting the motion of the large-scale (>1000 mi or 1600 km) pressure systems in the atmosphere for relatively short periods of time (1–5 days). For such space and time scales, the poorly understood energy sources and frictional dissipative forces may

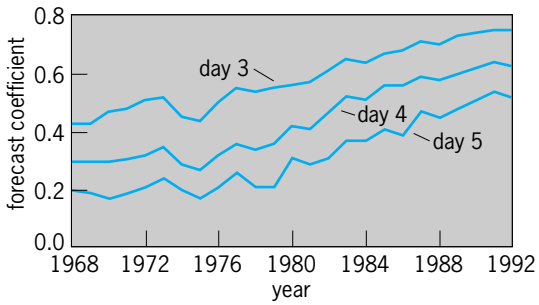


Fig. 1. Correlation coefficients between 3-day, 4-day, and 5-day forecasts of sea-level pressure made by the National Meteorological Center and observed sea-level pressure. Domain of verification is North America and adjacent ocean areas. (*Courtesy Wayman Baker, National Meteorological Center*)

be approximated by relatively simple formulations, and rather coarse horizontal resolutions (60–120 mi or 100–200 km) may be used.

Great progress has been made in improving the accuracy of numerical weather prediction models. Forecasts for 3, 4, and 5 days have steadily improved; the 5-day forecast is now as accurate as the 3-day forecast was in the early 1970s (**Fig. 1**). Because of the increasing accuracy of numerical forecasts, numerical models have become the basis for medium-range (1–10 days) forecasts made by the weather services of most countries. *See* MESOMETEOROLOGY.

Cloud and precipitation prediction. If, to the standard dynamic variables, the density of water vapor is added, it becomes possible to predict clouds and precipitation in addition to the air motion. When a parcel of air containing a fixed quantity of water vapor ascends, it expands adiabatically and cools until it becomes saturated. Continued ascent produces clouds and precipitation. The most successful predictions made by this method are obtained in regions of strong rising motion, whether induced by forced orographic ascent or by horizontal convergence in well-developed cyclones. The physics and mechanics of the convective cloud-formation process make the prediction of convective cloud and showery precipitation more difficult.

Global prediction. In 1955 the first operational numerical weather prediction model was introduced at the National Meteorological Center. This simplified barotropic model consisted of only one layer, and therefore it could model only the temporal variation of the mean vertical structure of the atmosphere. By the early 1990s, the speed of computers had increased sufficiently to permit the development of multilevel (usually about 10–20) models that could resolve the vertical variation of the wind, temperature, and moisture. These multilevel models predict the fundamental meteorological variables for large scales of motion. Global models with horizontal resolutions as fine as 125 mi (200 km) are being used by weather services in several countries.

Global numerical weather prediction models require the most powerful computers to complete a 10-day forecast in a reasonable amount of time. For example, a 10-day forecast with a 15-layer global model with horizontal resolution of 90 mi (150 km) requires approximately 10^{12} calculations. A supercomputer capable of performing 10^8 arithmetic operations per second would then require 10^4 s or 2.8 h to complete a single 10-day forecast.

Numerical models of climate. While global models were being implemented for operational weather prediction 1–10 days in advance, similar research models were being developed that could be applied for climate studies by running for much longer time periods. The extension of numerical predictions to long time intervals (many years) requires a more accurate numerical representation of the energy transfer and turbulent dissipative processes within the atmosphere and at the air-earth boundary, as well as greatly augmented computing-machine speeds and capacities.

With state-of-the-art computers, it is impossible to run global climate models with the same high resolution as numerical weather prediction models; for example, a hundred-year climate simulation with the numerical weather prediction model that requires 2.8 h for a 10-day forecast would require 10,220 h (1.2 years) of computer time. Therefore, climate models must be run at lower horizontal resolutions than numerical weather prediction models (typically 250 mi or 400 km).

Predictions of mean conditions over the large areas resolvable by climate models are feasible because it is possible to incorporate into the prediction equations estimates of the energy sources and sinks—estimates that may be inaccurate in detail but generally correct in the mean. Thus, long-term simulations of climate models with coarse horizontal resolutions have yielded simulations of mean circulations that strongly resemble those of the atmosphere. These simulations have been useful in explaining the principal features of the Earth's climate, even though it is impossible to predict the daily fluctuations of weather for extended periods. Climate models have also been used successfully to explain paleoclimatic variations, and are being applied to predict future changes in the climate induced by changes in the atmospheric composition or characteristics of the Earth's surface due to human activities. *See* CLIMATE HISTORY; CLIMATE MODIFICATION.

Limited-area models. Although the relatively coarse grids in global models are necessary for economical reasons, they are sources of two major types of forecast error. First, the truncation errors introduced when the continuous differential equations are replaced with approximations of finite resolution cause erroneous behavior of the scales of motion that are resolved by the models. Second, the neglect of scales of motion too small to be resolved by the mesh (for example, thunderstorms) may cause errors in the larger scales of motion. In an effort to simultaneously reduce both of these errors, models with considerably finer meshes have been tested. However, the price of reducing the mesh has been the necessity of covering smaller domains in order to keep the total computational effort within computer capability. Thus the main operational limited-area model run at the National Meteorological Center has a mesh length of

approximately 50 mi (80 km) on a side and covers a limited region approximately two times larger than North America. Because the side boundaries of this model lie in meteorologically active regions, the variables on the boundaries must be updated during the forecast. A typical procedure is to interpolate these required future values on the boundary from a coarse-mesh global model that is run first. Although this method is simple in concept, there are mathematical problems associated with it, including overspecification of some variables on the fine mesh. Nevertheless, limited-area models have made significant improvements in the accuracy of short-range numerical forecasts over the United States.

Even the small mesh sizes of the operational limited-area models are far too coarse to resolve the detailed structure of many important atmospheric phenomena, including hurricanes, thunderstorms, sea- and land-breeze circulations, mountain waves, and a variety of air-pollution phenomena. Considerable effort has gone into developing specialized research models with appropriate mesh sizes to study these and other small-scale systems. Thus, fully three-dimensional hurricane models with mesh sizes of 5 mi (8 km) simulate many of the features of real hurricanes. On even smaller scales, models with horizontal resolutions of a few hundred meters reproduce many of the observed features in the life cycle of thunderstorms and squall lines. It would be misleading, however, to imply that models of these phenomena differ from the large-scale models only in their resolution. In fact, physical processes that are negligible on large scales become important for some of the phenomena on smaller scales. For example, the drag of precipitation on the surrounding air is important in simulating thunderstorms, but not for modeling large scales of motion. Thus the details of precipitation processes, condensation, evaporation, freezing, and melting must be incorporated into realistic cloud models.

In another class of special models, chemical reactions between trace gases are considered. For example, in models of urban photochemical smog, predictive equations for the concentration of oxides of nitrogen, oxygen, ozone, and reactive hydrocarbons are solved. These equations contain transport and diffusion effects by the wind as well as reactions with solar radiation and other gases. Such air-chemistry models become far more complex than atmospheric models as the number of constituent gases and permitted reactions increases. *See* COMPUTER PROGRAMMING; DIGITAL COMPUTER; MODEL THEORY; SUPERCOMPUTER.

Richard A. Anthes

Modern weather analysis and forecasting procedures involve (1) the collection of global meteorological surface and upper-air observations in place (for example, ships of opportunity) and remotely sensed (for example, satellite sensors) instrumentation platforms; (2) the preparation of global surface and upper air pressure, temperature, moisture, and wind analyses at frequent time intervals based upon these observations; (3) the solution of a closed set of highly nonlinear equations governing atmospheric dynamical motions by numerical means using the most sophisticated computers to predict the future state of the global atmosphere from the initial conditions specified by the analyses in step 2; (4) the application of statistical procedures such as model output statistics (MOS) to the atmospheric simulations obtained from the dynamical models to predict a wide variety of weather elements of interest to potential users by objective means; and (5) the monitoring the intervention by experienced human forecasters at appropriate stages in this objective forecasting process to improve the quality and breadth of the forecast information communicated to a wide spectrum of users in the private and public sectors.

Data collection and analysis. Surface meteorological observations are routinely collected from a vast continental data network, with the majority of these observations obtained from the middle latitudes of both hemispheres. Commercial ships of opportunity, military vessels, and moored and drifting buoys provide similar in-place measurements from oceanic regions, although the data density is biased toward the principal global shipping lanes. Information on winds, pressure, temperature, and moisture throughout the troposphere and into the stratosphere is routinely collected from (1) balloon-borne instrumentation packages (radiosonde observations) and commercial and military aircraft which sample the free atmosphere directly; (2) ground-based remote-sensing instrumentation such as wind profilers (vertically pointing Doppler radars), the National Weather Service Doppler radar network, and lidars; and (3) special sensors deployed on board polar orbiting or geostationary satellites. The remotely sensed observations obtained from meteorological satellites have been especially helpful in providing crucial measurements of areally and vertically averaged temperature, moisture, and winds in data sparse (mostly oceanic) regions of the world. Such measurements are necessary to accommodate modern numerical weather prediction practices and to enable forecasters to continuously monitor global storm (such as hurricane) activity. *See* LIDAR; METEOROLOGICAL INSTRUMENTATION; RADAR METEOROLOGY.

At major operational weather prediction centers such as the National Center for Environmental Prediction (NCEP, formerly known as the National Meteorological Center or the European Centre for Medium Range Weather Forecasts, ECMWF), the global meteorological observations are routinely collected, quality-checked, and mapped for monitoring purposes by humans responsible for overseeing the forecast process. The basic observations collected by NCEP are also disseminated on-line by Internet Data Distribution (IDD) or in mapped form by digital facsimile (DIFAX) to prospective users in academia and the public and private sectors. At NCEP and ECMWF (and other centers) the global observational data stream is further machine-processed in order to prepare a full three-dimensional set of global surface and upper air analyses of selected meteorological fields at representative time periods. The typical horizontal and vertical resolution of the globally gridded analyses is about 60 mi (100 km) and 25–50 hectapascals (millibars), respectively. Preparation of these

analyses requires a first guess field from the previous numerical model forecast against which the updated observations are quality-checked and objectively analyzed to produce an updated global gridded set of meteorological analyses.

These updated analyses are modified as part of a numerical procedure designed to ensure that the gridded meteorological fields are dynamically consistent and suitable for direct computation in the new forecast cycle. This data assimilation, analysis, and initialization procedure, known as four-dimensional data assimilation, was at one time performed four times daily at the standard synoptic times of 0000, 0600, 1200, and 1800 UTC. The four-dimensional data assimilation is performed almost continuously, given that advanced observational technologies (such as wind profilers; automated surface observations; automated aircraft-measured temperature, moisture, and wind observations) have ensured the availability of observations at other than the standard synoptic times mentioned above. An example is the rapid update cycle (RUC) mesoscale analysis and forecast system used at NCEP to produce real-time surface and upper air analyses over the United States and vicinity every 3 h and short-range (out to 12 h) forecasts. The rapid update cycle system takes advantage of high-frequency data assimilation techniques to enable forecasters to monitor rapidly evolving mesoscale weather features.

Operational models. The forecast models used at NCEP and ECMWF and the other operational prediction centers are based upon the primitive equations that govern hydrodynamical and thermodynamical processes in the atmosphere. The closed set of equations consists of the three momentum equations (east-west, north-south, and the vertical direction), a thermodynamic equation, an equation of state, a continuity equation, and an equation describing the hydrological cycle. Physical processes such as the seasonal cycle in atmospheric radiation, solar and long-wave radiation, the diurnal heating cycle over land and water, surface heat, moisture and momentum fluxes, mixed-phase effects in clouds, latent heat release associated with stratiform and convective precipitation, and frictional effects are modeled explicitly or computed indirectly by means of parametrization techniques. A commonly used vertical coordinate in operational prediction models is the sigma coordinate, defined as the ratio of the pressure at any point in the atmosphere to the surface (station) pressure.

The NCEP operational global prediction model is known as the medium-range forecast (MRF) model. The medium-range forecast is run once daily out to 14 days from the 0000 UTC analysis and initialization cycle in support of medium-range and extended-range weather prediction activities at NCEP. The medium-range forecast run commences shortly before 0900 UTC in order to allow sufficient time for delayed 0000 UTC surface and upper air data to reach NCEP and to be included in the analysis and initialization cycle.

The aviation forecast model (AVN) is run from the 1200 UTC analysis and initialization cycle in support of NCEP's global aviation responsibilities. The aviation forecast model and medium-range forecast are identical models except that the AVN is run with a data cut off of about 3 h and only out to 72 h in order to minimize the time necessary to get the model forecast information to the field. The MRF runs with a vertical resolution of 38 levels, a horizontal resolution of about 63 mi (105 km) [triangular truncation in spectral space] for the first 8 days, and a degraded horizontal resolution of about 130 mi (210 km) for the last 6 days of the 14-day operational forecast.

The medium-range forecast is also the backbone of the ensemble forecasting effort (a series of parallel medium-range forecast runs with slightly changed initial conditions and medium-range forecast runs from different time periods) that was instituted in the early 1990s to provide forecaster guidance in support of medium-range forecasting (6–10-day) activities at NCEP. The scientific basis for ensemble prediction is that model forecasts for the medium and extended range should be considered stochastic rather than deterministic in nature in recognition of the very large forecast differences that can occur on these time scales between two model runs initialized with only very slight differences in initial conditions. Continuing advances in computer power have made it practical to begin operational ensemble numerical weather prediction. Given that each member of the ensemble (the ensemble consists of 38 different model runs) represents an equally likely model forecast outcome, the spread of the ensemble forecasts is taken as a measure of the potential skill of the forecasts. Forecasters look for clustering of the ensemble members around a particular solution as indicative of higher probability outcomes as part of a process inevitably known as "forecasting the forecasts". Ensemble forecasting techniques are also taking hold at other operational centers such as the ECMWF, and they are likely to spread to regional models such as the NCEP Eta and Regional Analysis and Forecast (RAFS) models in the nea future.

Regional forecast models such as the NCEP Eta model are utilized to help predict atmospheric circulation patterns associated with convective processes and terrain forcing. Given the increasing importance of forecasting significant mesoscale weather events (such as squall lines, flash flood occurrences, and heavy snow bands) that occur on time scales of a few hours and space scales of a few hundred kilometers, the Eta model will continue to be developed at NCEP in support of regional model guidance to local forecasters. An important challenge to mesoscale models is to make better precipitation forecasts, especially of significant convective weather events. Success in this endeavor will require increased use of satellite- and land-based data sets, especially measurements of water vapor, that are coming on line.

Forecast products and forecast skill. These are classified as longer term (greater than 2 weeks) and shorter term.

Longer term. The Climate Prediction Center (CPC) of NCEP initiated a long-lead climate outlook program for the United States in early 1995. The scientific basis for this initiative rests upon the understanding

of the coupled nature of global atmospheric and oceanic circulations on intraseasonal and interannual time scales. Particular attention has been focused on the climatic effects arising from low-frequency changes in global atmospheric wind, pressure, and moisture patterns associated with evolving tropical oceanic sea surface temperature and thermocline depth anomalies known as part of the El Niño–Southern Oscillation (ENSO) climate signal. Statistically significant temperature and moisture anomalies have been identified over portions of the conterminous United States (and in many other regions of the world) that can occur in association with an ENSO event, especially during the cooler half of the year. For example, in winter there is a tendency for the southern United States from Texas to the Carolinas, Georgia, and Florida to be wetter and cooler than normal, while much of the Pacific Northwest eastward across southern Canada and the northern United States to the Great Lakes is warmer and drier than normal during the warm phase (anomalies in sea surface temperature in the tropical Pacific from the Dateline eastward to South America are positive, while sea-level pressure anomalies are positive and negative, respectively, over the western versus central and eastern Pacific) of an ENSO event. *See* EL NIÑO; TROPICAL METEOROLOGY.

Similarly, the tropical convective activity that usually maximizes over the western Pacific oceanic warm pool (sea surface temperature about 86°F or 30°C) tends to shift eastward toward the central and, occasionally, eastern Pacific during the warm phase of an ENSO event. Under such conditions, drought and heat-wave conditions may prevail over portions of Australia and Indonesia, while conditions become much wetter in the normally arid coastal areas along western Peru and northern Chile. During the exceptionally strong ENSO event of 1982–1983, extreme climatic anomalies were observed at many tropical and midlatitude locations around the world. During the cold phase of ENSO (known as La Niña) when anomalies in the sea surface temperature in the tropical eastern Pacific are negative and the easterly trade winds are stronger than average, many of these global climatic anomalies may reverse sign, such as occurred during the 1988–1989 cold event. Increased knowledge and understanding of the global aspects of atmospheric and oceanic low-frequency phenomena have resulted from combined observational, theoretical, and numerical research investigations that in part were stimulated by the extraordinary warm ENSO event of 1982–1983. The knowledge is now being applied toward the preparation of long-lead seasonal forecasts at the Climate Prediction Center and elsewhere. *See* DROUGHT.

In the extended forecasting procedure of the Climate Prediction Center, a total of 26 probability anomaly maps are prepared (one each for temperature and precipitation for equally likely below-normal and above-normal categories) for each of 13 lead times from 0.5 to 12.5 months. The probabilistic format (begun in 1982 with the issuance of monthly outlooks of temperature and precipitation for the United States) and accompanying skill assessment of these long-lead forecasts allows sophisticated users in the public and private sectors to make better use of the outlooks for a variety of business and planning purposes. The long-lead climate outlooks are made from a combination of the following methods: (1) a statistical procedure known as canonical correlation analysis that relates prior anomalies of global sea surface temperatures, Northern Hemisphere 700-hPa (mbar) heights, and temperature and precipitation anomalies over the United States to subsequent anomalies; (2) another statistical procedure known as optimal climate normals that projects the persistence of recent temperature and precipitation anomalies to future months and seasons; and (3) forecasts of temperature and precipitation anomalies obtained from a coupled atmosphere-ocean general circulation model integrated forward in time 6 months by using selected sea surface temperature anomalies.

Shorter term. An important aspect of the forecasting process is the interpretation and dissemination of model forecasts to a wide variety of users and the general public. Specialized NCEP centers (in addition to the Climate Prediction Center) such as the National Severe Storms Forecasting Center (responsible for issuing watches and warnings of severe weather events), the Storm Prediction Center (charged with monitoring a broad range of winter and summer storms), the National Hurricane Center (responsible for monitoring the hurricane threat), and the Hydrological Meteorological Center (responsible for monitoring river basin water storage, runoff, and flood potential) are charged with providing guidance to local National Weather Service forecast offices about a variety of weather hazards. Experienced meteorologists and hydrologists at these centers are responsible for the interpretation of NCEP model and statistical forecasts so as to provide additional guidance and information on upcoming important weather events to individual offices of the National Weather Service. As an example, the Heavy Precipitation Branch of the NCEP has the responsibility of providing forecasts of significant rain and snow amount areas over the continental United States out to 48 h in advance. The experienced forecasters in the Heavy Precipitation Branch have been able to maintain their skill over the model objective precipitation forecasts since the mid-1970s, despite demonstrated model forecasting improvements, through a combination of detailed knowledge of regional and local mesometeorology, an understanding of systematic model biases and how to correct for them, and a continuing ability to select the best model on any given day in a form of human ensemble forecasting.

The skill of numerical weather prediction forecasts of temperature and winds in the free atmosphere [for example, the 500 hPa (mbar) level] continues to improve steadily. For example, since the mid-1980s the root-mean-square vector error of the 24-h 250-hPa (mbar) wind forecasts produced by the operational global NCEP model has declined from 20 mi/h to 16 mi/h (10 m s^{-1} to 7.5 m s^{-1}). Experience with medium-range-forecast/ECMWF day five predictions of mean sea-level

pressure [500 hPa (mbar) heights] over North America indicates that standardized anomaly correlation coefficient scores (a standardized anomaly is the anomaly divided by the climatological standard deviation and is used to avoid overweighting anomalies at higher latitudes relative to lower latitudes) range from 30 to 65 (experience suggests a standardized score >20 has operational forecast value) with the higher (lower) scores occurring in winter (summer). At NCEP the human forecasters have been able to improve upon the medium-range-forecast anomaly correlation coefficient scores for day five mean sea-level pressure forecasts.

Forecasts of surface weather elements such as temperature, precipitation type and amount, ceiling, and visibility are prepared objectively by a statistical-dynamical model output statistics (MOS) technique. The cornerstone of the MOS approach is the preparation of a series of regression equations for individual weather elements at specified locations, with the regression coefficients determined from predictors provided by operational NCEP models. MOS guidance is relayed to National Weather Service offices and a wide variety of external users. The MOS product has become very competitive with the human forecaster as measured by standard skill scores. Over the 25-year period ending in 1992, the mean absolute error of National Weather Service maximum and minimum temperatures issued to the general public decreased by about 1.08°F (0.6°C). To put this number in perspective, the skill of 36-h temperature forecasts issued in the late 1990s is comparable to the skill of similar 12-h forecasts made in the late 1960s. A similar "24-h gain" in skill level has been registered for routine 12-h probability-of-precipitation forecasts of the National Weather Service. Skill levels are a function of the forecast type and projection. Skill (as measured relative to a climatological control by using a standard skill measure such as the Brier score) in day-to-day temperature forecasting approaches zero by 7–8 days in advance. For probability-of-precipitation forecasts the zero skill level is typically reached 2–3 days earlier. Precipitation-amount forecasts typically show little skill beyond 3 days, while probability-of-thunderstorm forecasts may be skillful only out to 1–2 days ahead.

These varying skill levels reflect the fact that existing numerical prediction models such as the medium-range forecast have become very good at making large-scale circulation and temperature forecasts, but are less successful in making weather forecasts. An example is the prediction of precipitation amount and type given the occurrence of precipitation and convection. Each of these forecasts is progressively more difficult because of the increasing importance of mesoscale processes to the overall skill of the forecast. *See* PRECIPITATION (METEOROLOGY).

Lance F. Bosart

Nowcasting

Nowcasting is a form of very short range weather forecasting. The term nowcasting is sometimes used loosely to refer to any area-specific forecast for the period up to 12 h ahead that is based on very detailed observational data. However, nowcasting should probably be defined more restrictively as the detailed description of the current weather along with forecasts obtained by extrapolation up to about 2 h ahead. Useful extrapolation forecasts can be obtained for longer periods in many situations, but in some weather situations the accuracy of extrapolation forecasts diminishes quickly with time as a result of the development or decay of the weather systems.

Comparison with traditional forecasting. Much weather forecasting is based on the widely spaced observations of temperature, humidity, and wind obtained from a worldwide network of balloon-borne radiosondes. These data are used as input to numerical-dynamical weather prediction models in which the equations of motion, mass continuity, and thermodynamics are solved for large portions of the atmosphere. The resulting forecasts are general in nature. Although these general forecasts for one or more days ahead have improved in line with continuing developments of the numerical-dynamical models, there have not been corresponding improvements in local forecasts for the period up to 12 h ahead. The problem is that most of the mathematical models cope adequately with only the large weather systems, such as cyclones and anticyclones, referred to by meteorologists as synoptic-scale weather systems. *See* WEATHER.

Nowcasting involves the use of very detailed and frequent meteorological observations, especially remote-sensing observations, to provide a precise description of the "now" situation from which very short-range forecasts can be obtained by extrapolation. Of particular value are the patterns of cloud, temperature, and humidity which can be obtained from geostationary satellites, and the fields of rainfall and wind measured by networks of ground-based radars. These kinds of observations enable the weather forecaster to keep track of smaller-scale events such as squall lines, fronts and thunderstorm clusters, and various terrain-induced phenomena such as land-sea breezes and mountain-valley winds. Meteorologists refer to these systems as mesoscale weather systems because their scale is intermediate between the large or synoptic-scale cyclones and the very small or microscale features such as boundary-layer turbulence. *See* FRONT; METEOROLOGICAL SATELLITES; SQUALL LINE; STORM.

There is a distinction between traditional forecasting and nowcasting in terms of the lead time and quality of the forecast (**Fig. 2**). The two approaches are complementary; each has its place according to the lead time and detail required of the forecast. Conceptually, nowcasting is a simple procedure. However, vast amounts of data are involved, and it is only since the development of the necessary digital data-processing, transmission, and display facilities that it has been possible for nowcasting to be available economically. *See* DATA COMMUNICATIONS.

Information technology. A key problem in nowcasting is that of combining diverse and complex data streams, especially conventional meteorological data with remote-sensing data. It is widely held that this combination should be achieved by using digital

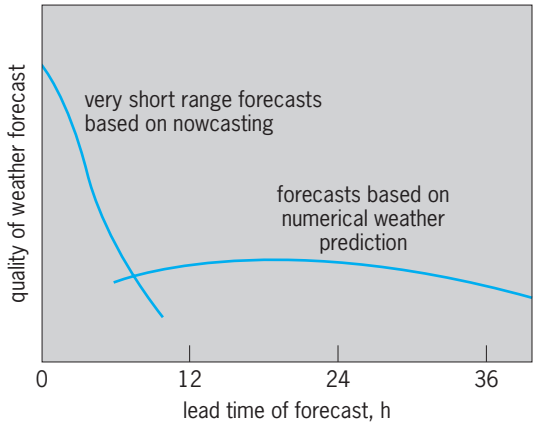


Fig. 2. The quality of weather forecasts, defined as the product of the accuracy and detail achievable, depicted schematically as a function of lead time for two different forecasting methods. Traditional forecasts based on numerical weather prediction deal with synoptic-scale events such as cyclones and anticyclones; very short range forecasts based on the nowcasting approach resolve so-called mesoscale events, such as thunderstorms, fronts, and squall lines.

data sets displayed on interactive video displays—the so-called workstation concept. By means of advanced human-computer interaction techniques, the weather forecaster will eventually be able to analyze the merged data sets by using a light pen or a finger on touch-sensitive television screens. Various automatic procedures can be implemented to help the forecaster carry out analyses and extrapolation forecasts, but the incomplete nature of the data sets is such that the forecaster will almost always be in the position of needing to fine-tune the products subjectively. The idea behind the forecasting workstation is to simplify the routine chores of basic data manipulation so that the forecaster is given the maximum opportunity to exercise judgment within the context of what is otherwise a highly automated system.

Very short-range forecast products are highly perishable: they must be disseminated promptly if they are not to lose their value. Advances in technology offer means for the rapid tailoring and dissemination of the digital forecast information and for presenting the material in convenient customer-oriented formats. In some cases, dissemination may be by direct computer-to-computer link with the user's control system. In another cases, the user will benefit from methods of visual presentation using such media as cable television or viewdata.

Nowcasting and mesoscale models. There is some evidence of a gap in forecasting capability between nowcasting and traditional synoptic-scale numerical weather prediction. This occurs for forecast lead times between about 6 and 12 h, that is, for periods when development and decay are beginning to invalidate forecasts by simple extrapolation (Fig. 2). To some extent the forecaster can identify some of the likely developments by interpreting the nowcast information in the light of local climatologies and conceptual life-cycle models of weather systems. But the best way to forecast changes is to use numerical-dynamical methods. It would be natural to assume that an immediate way would be to incorporate the detailed (but usually incomplete) nowcast data as input to numerical-dynamical models with a finer resolution than those presently in operational use; these are the so-called mesoscale numerical models. Research into the use of detailed observational data in mesoscale models has been actively pursued, but significant technical difficulties have slowed progress.

Keith A. Browning

Extended-Range Prediction

Forecasts of time averages of atmospheric variables, for example, sea surface temperature, where the lead time for the prediction is more than 2 weeks, are termed long-range or extended-range climate predictions. The National Weather Service issues extended-range predictions of monthly and seasonal average temperature and precipitation near the middle of each month; these are known as climate outlooks. Monthly outlooks have a single lead time of 2 weeks. The seasonal outlooks have lead times ranging from 2 weeks to 12 $^1/_2$ months in 1-month increments. Thus, in each month a set consisting of 1 monthly and 13 seasonal long-lead-time climate outlooks is released.

Climate versus weather. When the daily average temperature (that is, the average of the high and the low temperatures for each day) is plotted as a graph for a given point, the result is a wildly oscillating curve (**Fig. 3**). Such a curve represents weather, the details of which are considered to be unpredictable beyond about 2 weeks. A curve for the 30-year average of the daily average temperature represents the climatology (Fig. 3). Finally, a 90-day running average can be applied to the curve that represents the daily average (the heaviest curve in Fig. 3); this represents the short-term seasonal climate behavior. The differences between this curve and the climatology represent the kind of temperature and, by analogy, the pressure predictions discussed in this section.

Scientific basis. Prior to the early 1980s, the techniques used to produce extended-range predictions assumed that the 700-mbar height field integrates the effects of long-lived atmospheric processes, which produce the observed 700-mbar anomalies. These signals can sometimes be isolated via linear statistical techniques, even though in general the physical causes of the anomalies are unknown. Predictions made by these techniques were slightly more accurate than climatology. During the 1980s, it was discovered that the ENSO produces, by far, the largest measurable short-term variability of middle-latitude climate of any low-frequency phenomenon known.

The name El Niño applies to the warm phase of the Southern Oscillation. On average, every 4–5 years the sea surface temperature in the central and eastern equatorial Pacific Ocean warms 1°C or more above its 30-year average. This is accompanied by weakening of the Pacific trade winds, a decrease of sea-level pressures near Tahiti in the eastern tropical Pacific, an increase of sea-level pressures in northern Australia, an increase in the depth of the oceanic thermocline in the tropical eastern Pacific Ocean (that is, anomalously warm water replaces climatologically cold water), and a shift of deep cumulus convection and rainfall from the western equatorial Pacific into

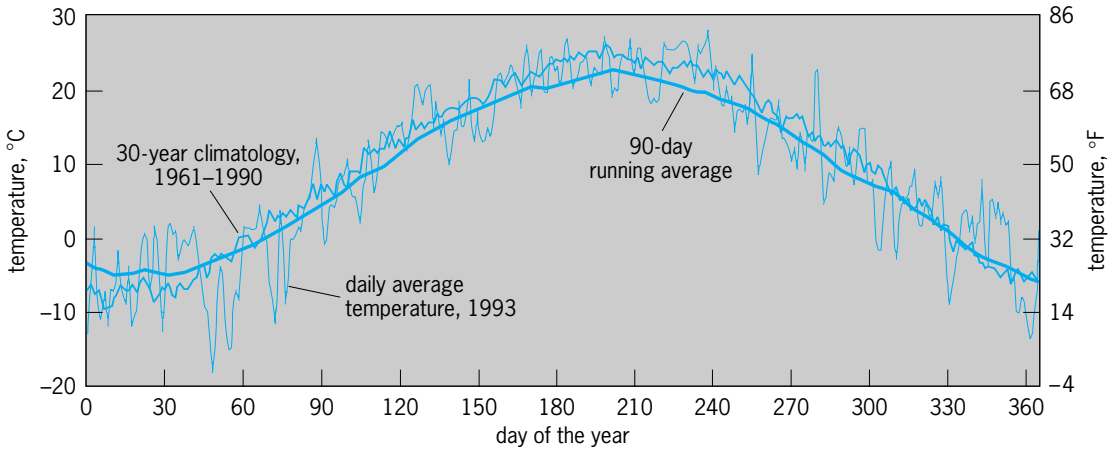


Fig. 3. Daily average temperature at Des Moines, Iowa. The thinnest curve indicates daily average temperature for January 1–December 31, 1994. The thickest curve indicates the 90-day running mean. The remaining curve indicates the 30-year average of daily average temperature, 1961–1990.

the central and eastern Pacific. The shift in deep convection alters the global atmospheric circulation, perturbing the climatological transports of heat and momentum. The warm phase often reaches this mature state by late December and normally lasts up to 18 months. This pattern is sometimes followed by a transition to the cold phase of the Southern Oscillation, although the periodicity is not particularly regular; for example, warm-phase conditions have on occasion lasted for three or more consecutive years (1939–1941, 1991–1995).

The anomalous tropospheric warming in the eastern tropical Pacific Ocean during the warm phase strengthens the Hadley circulation, including the subtropical jet. This energized subtropical jet contributes directly to a greater frequency and strength of winter and springtime storms over the southern United States, the effect sometimes including California, but it almost always brings enhanced precipitation to Texas and the Gulf states. The polar jet is also affected, often being constrained to lie north of its normal position over North America. This leads to warmer and drier conditions over sections of the northern United States. Opposite conditions tend to prevail during the cold phase of the Southern Oscillation, although the amplitude of the surface anomalies tends to be smaller than during the warm phase.

Models. The long life span and pronounced effect that ENSO has upon short-term climate variability, over the United States and elsewhere, clearly presents an opportunity to make predictions based upon the status of ENSO with a lead time of as much as a year. The statistical models traditionally used in long-range prediction have been reformulated to take advantage of this knowledge. This enhanced physical understanding has permitted, for the first time, the development and use in operational seasonal prediction of a physical model like those used for numerical weather prediction, which couples the ocean and the atmosphere. Such models should improve seasonal prediction. *See* ATMOSPHERIC GENERAL CIRCULATION; JET STREAM; TROPOSPHERE.

Physically based coupled model predictions of seasonal averages are achieved at the National Weather Service by using a global spectral medium-range forecast model modified to have improved response to sea surface temperature forcing the atmosphere in the tropics. Two 6-month runs of the model, driven by two different sets of sea surface temperature, are made. The first of the runs uses sea surface temperature persisting from a half month ago, while the second sea surface temperature fields are predicted by a three-member ensemble of coupled ocean/atmosphere model runs.

Three examples of objective empirical prediction methods are mixed persistence analog, canonical correlation analysis, and optimal climate normals. The mixed persistence analog consists of first determining, from 45 years of observations, the 10 years that are best analogs to one or more of the most recently observed seasonal means of United States surface temperature, the sea surface temperature, the average temperature of the lower 1.8 mi (3 km) of the atmosphere, and the seasonal mean 700-hPa (mbar) height-anomaly pattern. The mix of predictors used is based upon the results of forecasts made in prior years and varies with the season. A final analog is formed from a similarity-based weighted composite of these 10 individual analog years. The temperature pattern represented by this final analog is combined, via multiple linear regression, with the most recently observed seasonal mean temperature anomaly pattern (that is, persistence).

Canonical correlation analysis is a multivariate statistical technique that relates future spatial and temporal patterns of United States surface temperature or pressure to spatial and temporal patterns of observed seasonal mean tropical Pacific sea surface temperature, hemispheric 700-hPa (mbar) height, and United States surface temperature. Thus, predictor spatial patterns at different prior times can be related to predict and spatial patterns at different future times.

The optimal climate normals method is based upon the simple notion that it is possible to obtain a better estimate of the average state of the climate (seasonal mean temperature or pressure) for a future season than the standard 30-year climatology

by computing, for that season, the average of the most recent k years, where k is less than 30 years. For operational purposes the value of k is fixed at 10 years for temperature and 15 years for pressure. The optimal climate normals method isolates effects of trends or recent multiyear regimes in seasonal observed time series.

Accuracy. The accuracy of long-range outlooks has always been modest because the predictions must encompass a large number of possible outcomes, while the observed single event against which the outlook is verified includes the noise created by the specific synoptic disturbances that actually occur and that are unpredictable on monthly and seasonal time scales. According to some estimates of potential predictability, the noise is generally larger than the signal in middle latitudes.

An empirical verification method called cross-validation is used to evaluate most statistical models. In cross-validation the model being evaluated is formulated by using all observations except for one or more years, one of which is then predicted by the model. This is done, in turn, for all the years for which observations are available. The observations for each of the years withheld are then compared with their associated outlooks and the forecast accuracy is determined.

For coupled model forecasts, cross-validation is not used. Rather, the model must be rerun and verified for at least 10 years (cases) for a given season. These coupled model reruns, based upon observed sea surface temperatures, indicate that these forecasts of temperature have a cold-season United States average maximum correlation with observations of about 0.50 and a warm-season minimum score near zero. The average maximum score is reduced by about half when strong ENSO years are excluded. Coupled model precipitation forecasts over the United States have early winter and early spring accuracy maxima (correlations with observations of 0.30–0.40) when ENSO cases are included. Correlations between forecast and observed precipitation fields for this parameter also decline to about 0.10–0.15 when strong ENSO cases are excluded.

When forecasts of seasonal mean temperature made using the canonical correlation technique are correlated with observations, accuracy maxima in winter and summer and minima in spring and autumn are obtained. The score at lead times of 1, 4, and 7 months is nearly the same. Only at lead times of 10 and 13 months does the correlation drop to about half its value at the shorter lead times. As measured by the correlation between forecasts and observations, temperature forecasts made by the optimal climate normals technique have summer and winter maxima and minima in late winter–early spring and autumn. One interesting consequence of these temperature outlook accuracy properties is that, at certain times of the year, short-lead predictions may have lower expected accuracy than much longer lead predictions.

The accuracy of precipitation outlooks is generally lower than that of temperature outlooks. Precipitation predictions of the canonical correlation analysis and the optimal climate normals techniques have wintertime maximum correlations with observations. However, the former has low summertime accuracy, while the latter has relatively high accuracy during that season. For both methods, areas of high expected accuracy on the precipitation maps are much smaller and more spotty than those for temperature. This is due to the fact that precipitation is much more noisy (that is, variable on small space and time scales) than temperature.

Forecasts of seasonal mean temperature made by canonical correlation analysis and optimal climate normals have average correlations with observations in excess of 0.5 over large portions of the eastern United States during January through March and over the Great Basin during July through September.

Formulating an outlook. Forecasts of monthly and seasonal mean surface temperature and total precipitation, for the conterminous United States, Alaska, and Hawaii, are derived from a quasiobjective combination of canonical correlation analysis, optimal climate normals, and the coupled ocean-atmosphere model. For each method, maps consisting of standardized forecast values at stations are produced by computer. The physical size of the plotted numbers indicates the accuracy of the model. Two forecasters examine the map or maps for each lead time and variable, and resolve any conflicts among the methods. The statistical confidence in the resulting forecast will be low in regions where two or more equally reliable tools are in opposition or where the expected accuracy is low for all methods. Confidence is relatively high when tools agree and when the expected accuracy is high.

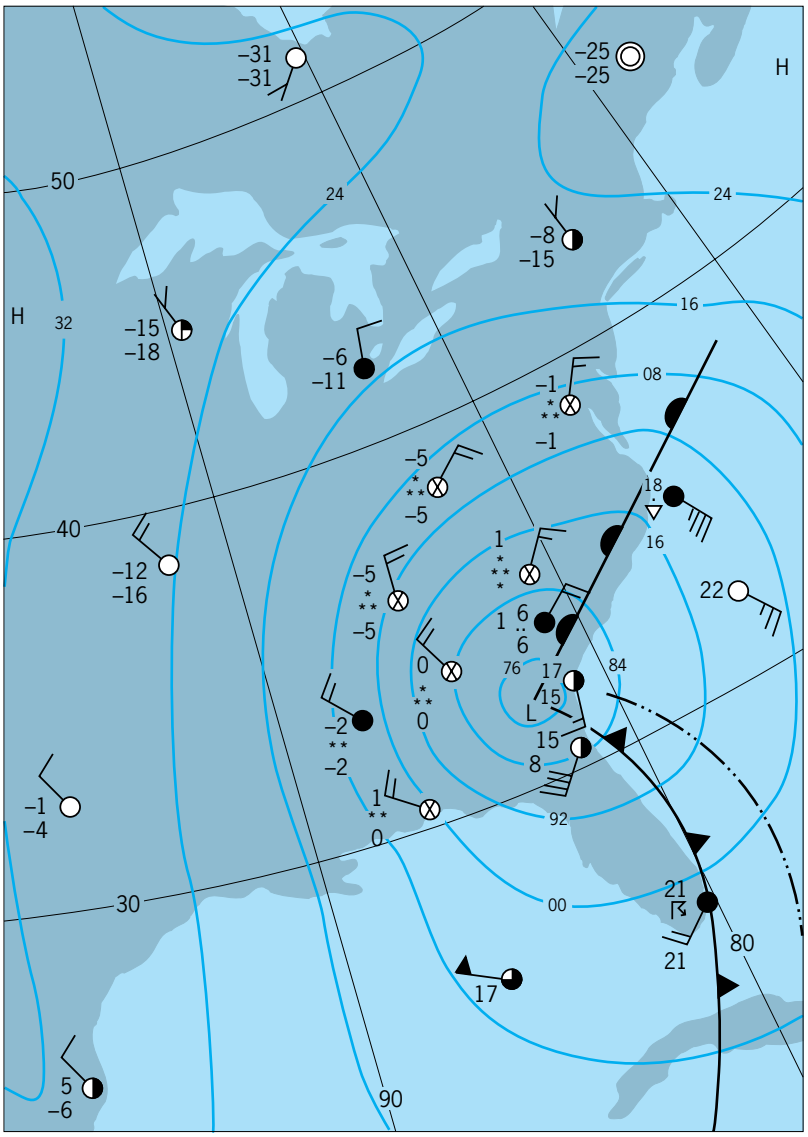
Information about accuracy and confidence is quantified through the use of probability anomalies to indicate the degree of confidence available from the tools. The climatological probabilities, with respect to which the outlook anomalies deviate, are determined well ahead of time. This is done by first dividing the observations of seasonal mean temperature and total precipitation during 1961–1990 into three equally likely categories for each of the stations at which forecasts are made. These are BELOW, NORMAL, and ABOVE for temperature, and BELOW, MODERATE, and ABOVE for precipitation. The likelihood of these categories is $^1/_3$-$^1/_3$-$^1/_3$ and is designated CL (for climatological probabilities) on the maps. Regions in which the correlation between the forecasts made by the tools and the observations, as determined by cross-validation, is less than 0.30 are given this designation. Elsewhere, nonzero probability anomalies—the numerical size of which is based upon the correlation scores of the tools, the degree of agreement among the tools, and the amplitude of the composite predicted anomalies—are assigned. Finally, the forecasters sketch the probability anomaly maps. *See* WEATHER MAP. Edward A. O'Lenic

Bibliography. C. D. Ahrens, *Meteorology Today With Infotrac: An Introduction to Weather, Climate, and the Environment*, 6th ed., 1999; K. A. Browning (ed.), *Nowcasting*, 1982; R. J. Doviak and D. S. Zrnic, *Doppler Radar & Weather Observations*, 2d ed., 1993; E. W. Friday, Jr., The modernization and associated restructuring of the National Weather

Service: An overview, *Bull. Amer. Meteorol. Soc.*, 75:43–52, 1994; J. R. Holton, *An Introduction to Dynamic Meteorology*, 3d ed., 1992; F. K. Lutgens, E. J. Tarbuck, and D. Tasa, *The Atmosphere: An Introduction to Meteorology*, 9th ed., 2003; F. Nebeker, *Calculating the Weather: Meteorology in the 20th Century*, 1995; P. Santurette and C. Georgiev, *Weather Analysis and Forecasting: Applying Satellite Water Vapor Imagery and Potential Vorticity Analysis*, 2005; S. Tracton and E. Kalnay, Operational ensemble prediction at the National Meteorological Center: Practical aspects, *Weath. Forecast.*, 8:379–398, 1993.

Weather map

A map or a series of maps that is used to depict the evolution and life cycle of atmospheric phenomena at selected times at the surface and in the free atmosphere. Weather maps are used for the analysis and display of in-place observational measurements and computer-generated analysis and forecast fields derived from weather and climate prediction models by research and operational meteorologists, government research laboratories, and commercial firms. Similar analyses derived from sophisticated computer forecast models are displayed in map form for forecast periods of 10–14 days in advance to provide guidance for human weather forecasters. *See* METEOROLOGICAL INSTRUMENTATION; WEATHER OBSERVATIONS.



Sea-level map illustrating isobars (solid lines every 8 hectopascals or 8 millibars) for 1200 UTC 13 March 1993. Plotted numbers adjacent to each station denote temperature (top) and dewpoint temperature (bottom) in ○C. Present weather is shown by conventional symbols. Wind speed is measured in m s^{-1} (1 pennant = 25 m s^{-1}; full barb = 5 m s^{-1}; half barb = 2.5 m s^{-1}). Bold lines and dash–double dot line denote conventional fronts and squall line position.

A sea-level weather map was prepared for the "Storm of the Century" at 1200 UTC (7 a.m. EST) 13 March 1993 (see **illus.**). At this time an intense 973-hectopascal (973-millibar) surface. cyclone was centered over southeastern Georgia. Heavy snow was falling from the Florida panhandle to southern New York, while severe weather, including heavy thunderstorms with hail and tornadoes, was reported across many portions of the Florida peninsula equatorward across Cuba and into Central America. The wind pattern at 500 hPa (500 millibar) was especially favorable for a major storm along the coast of eastern North America as the surface cyclone was situated beneath a strengthening southerly flow between a deepening trough over the Gulf of Mexico and an intensifying ridge over New England and the Canadian Maritime Provinces. Winds in the jet stream near 300 hPa (300 millibar) over the eastern United States exceeded 175 knots (90 m s^{-1}) as the storm intensified rapidly.

Rapid advances in computer technology and visualization techniques, as well as the continued explosive growth of the Internet distribution of global weather observations, satellite and radar imagery, and model analysis and forecast fields, have revolutionized how weather, climate, and forecast data and information can be conveyed to both the general public and sophisticated users in the public and commercial sectors. People and organizations with access to the Internet can access weather and climate information in a variety of digital or map forms in support of a wind range of professional and personal activities. *See* CLIMATOLOGY; METEOROLOGICAL SATELLITES; METEOROLOGY; RADAR METEOROLOGY; WEATHER FORECASTING AND PREDICTION.

Lance F. Bosart

Bibliography. C. D. Ahrens, *Meteorology Today: An Introduction to Weather, Climate and the Environment*, 6th ed., 1999; J. Williams, *The USA Today Weather Almanac*, 1995.

Weather modification

Human influence on the weather and, ultimately, climate. This can be either intentional, as with cloud seeding to clear fog from airports or to increase precipitation, or unintentional, as with air pollution, which increases aerosol concentrations and reduces sunlight. Weather is considered to be the day-to-day variations of the environment—temperature, cloudiness, relative humidity, wind-speed, visibility, and

precipitation. Climate, on the other hand, reflects the average and extremes of these variables, changing on a seasonal basis. Weather change may lead to climate change, which is assessed over a period of years. *See* CLIMATE HISTORY.

Specific processes of weather modification are as follows: (1) Change of precipitation intensity and distribution result from changes in the colloidal stability of clouds. For example, seeding of supercooled water clouds with dry ice (solid carbon dioxide, CO_2) or silver iodide (AgI) leads to ice crystal growth and fall-out; layer clouds may dissipate, convective clouds may grow. (2) Radiation change results from changes of aerosol or clouds (deliberately with a smoke screen, or unintentionally with air pollution from combustion), from changes in the gaseous constituents of the atmosphere (as with carbon dioxide from fossil fuel combustion), and from changes in the ability of surfaces to reflect or scatter back sunlight (as replacing farmland by houses.) (3) Change of wind regime results from change in surface roughness and heat input, for example, replacing forests with farmland.

Ice phase and cloud seeding. Water, when present in clouds in the atmosphere as droplets about 10 micrometers in diameter, often supercools—that is, exists as a metastable liquid—to temperatures as low as −40°F (−40°C). Random motion of water molecules in the liquid leads to the formation of clusters with an ice configuration; below −40°F a cluster can quickly grow to freeze the whole droplet (homogeneous nucleation). The vapor pressure over supercooled water is greater than over ice at the same temperature (**Fig. 1**), leading to preferential growth of ice crystals. A small ice crystal introduced into such a supercooled cloud grows to become a visible snow crystal about 1 mm in diameter in a few minutes, and is sufficiently large to fall out as precipitation (**Fig. 2**). Under natural conditions, ice crystals are only rarely found in growing clouds with temperature entirely above 14°F (−10°C). Below this temperature, ice crystals are found with increasing frequency as the temperature lowers. Here insoluble impurities (minerals, silver iodide, bacteria) nucleate ice at temperatures as high as 14 to 23°F (−10 to −5°C) and with increasing frequency as the temperature lowers to −40°F.

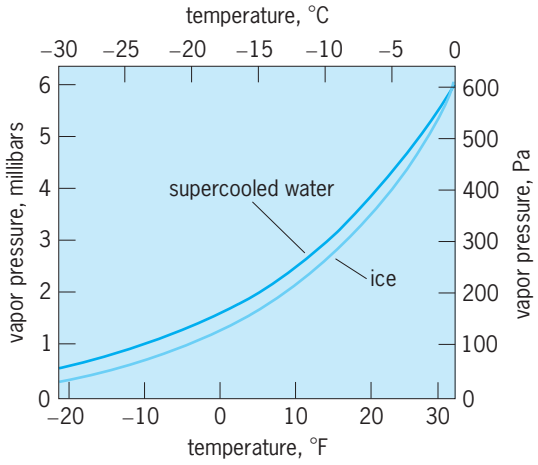


Fig. 1. Change with temperature of the vapor pressure of water vapor in equilibrium with supercooled water and ice.

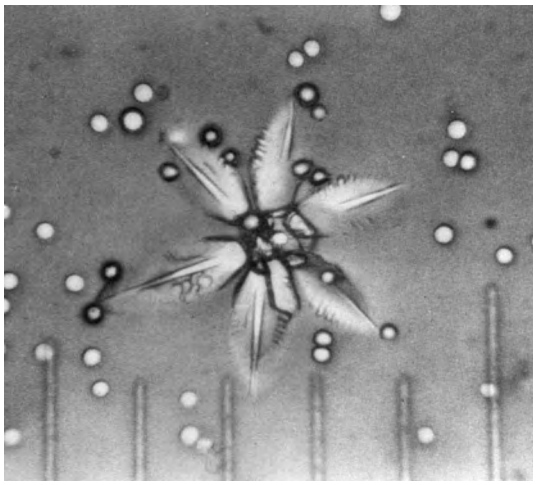


Fig. 2. Replica of cloud droplets and an ice crystal made during an aircraft penetration of a convective cloud in Montana. Each scale division represents 100 micrometers.

Aircraft measurements show a wide variation of ice crystal concentrations at a given temperature in different cloud types. Lenticular wave clouds formed near mountains sometimes are ice free at −31°F (−35°C); convective clouds over the ocean sometimes contain more than one crystal per liter of cloudy air just below 14°F (−10°C). Within this temperature range, ice crystals are nucleated on solid impurities of about 0.1 μm, having an atomic structure resembling ice (heterogeneous nucleation) carried upward from the Earth's surface—usually minerals such as kaolinite (clay) or possibly organic materials such as bacteria resulting from leaf decay.

Supercooled clouds may be induced to snow artificially by introduction of ice crystals in sufficient concentrations. This can be achieved in two ways. In the first method, cloudy air is cooled locally by dry-ice pellets at −108°F (−78°C) dropped from an aircraft. As they fall, they cool air in their wake below −40°F (−40°C), where droplets form and freeze. One kilogram of carbon dioxide produces 10^{14} crystals. A similar effect may be achieved by evaporation of liquid propane. The second method involves generating an impurity aerosol (diameter of about 0.05 μm) whose crystallographic structure is similar to ice, such as silver iodide or metaldehyde, and dispersing it into the cloud to form ice crystals [1 g AgI gives about 10^{15} (23°F or −5°C) to 10^{17} (5°F or −15°C) ice crystals].

Such an aerosol may be made by combustion of AgI-acetone solutions with a complexing additive. The aerosol generated at the ground is carried aloft under convective or turbulent conditions. Otherwise aircraft may carry continuous burners into the cloud, or drop flares containing silver compounds such as silver iodate ($AgIO_3$) which transforms to silver iodide (AgI) following combustion. Chlorine-containing compounds may be added to provide a silver iodate–silver chloride (AgI–AgCl) complex which gives a tenfold enhancement of ice nuclei.

Such flares may be dropped above cloud top, to burn as they fall through the cloud. The design of the pyrotechnic must be such that the smoke fails to coagulate prior to dilution, or the number of nuclei is dramatically reduced. Rockets have also been used to inject an explosive charge containing silver iodide into the cloud.

Historically, several projects employing these techniques have claimed an enhancement of precipitation by 10–15%. Assessment has been based on statistical evaluation of snowfall or rainfall at the ground. Since rainfall (along with most other meteorological quantities) is a highly variable quantity, and the effects produced are usually not large, such statistical tests need to be carried out over many seasons. It is necessary to evaluate whether any apparent rainfall enhancement as a departure from a mean is to be attributed to such variability or to the seeding process itself. Two techniques have been applied: use of an unseeded control area and cross-correlation of measured rainfall, and randomized seeding of one area and correlation of rainfall on seeded and nonseeded days. Each technique has problems of interpretation, since weather patterns producing precipitation change dramatically from day to day and from one year to the next. It might seem desirable to acquire for seeding trials as long a data record as practicable; however, too long a data record may lead to complications because of the possibility of changes on a climatic time scale unrelated to the changes being sought.

Physical evaluation of the seeding effectiveness is vital in removing uncertainties of an inadequate statistical base and the wide variability of a small number of precipitation situations. Aircraft penetration of seeded clouds reveals the presence of regions of supercooled water, a prerequisite, and whether or not the seeding aerosol has reached these regions to produce ice crystals in sufficient numbers (about one crystal per liter of air) to give measurable precipitation. Under ideal conditions, rainfall from the melting of snow has been measured by radar techniques and supplemented by rain gage measurements at the ground.

The most striking effects from seeding are observed in a stratiform cloud layer, or a layer of fog at the ground; dry-ice seeding from an aircraft typically produces an ice-crystal track some 0.3 mi (0.5 km) wide (**Fig. 3**). This procedure is used for clearing airports of overnight fog; it works well with cloud temperatures below about 25°F (−4°C). It has been suggested that this technique be used for clearing clouds over an urban area to allow sunlight to melt snow on the streets. By contrast, seeding cumulus clouds produces a visible result only rarely. In 1947, workers in Australia succeeded in causing a single cumulus cloud in a large number of such clouds to grow by dry-ice seeding. This was an example of dynamical seeding, resulting when release of latent heat by growth of crystals and freezing of droplets is sufficient to cause additional cloud convection. This is more likely when the cloud contains large quantities of supercooled water (2–4 $g \cdot m^{-3}$), particularly in the form of supercooled raindrops. An effect is more likely when such seeding is timed to enhance the maximum natural updraft. It has been suggested that seeding clusters of clouds leads to an overall increase in rainfall, since more moist air is incorporated into the more vigorously growing region than would occur naturally.

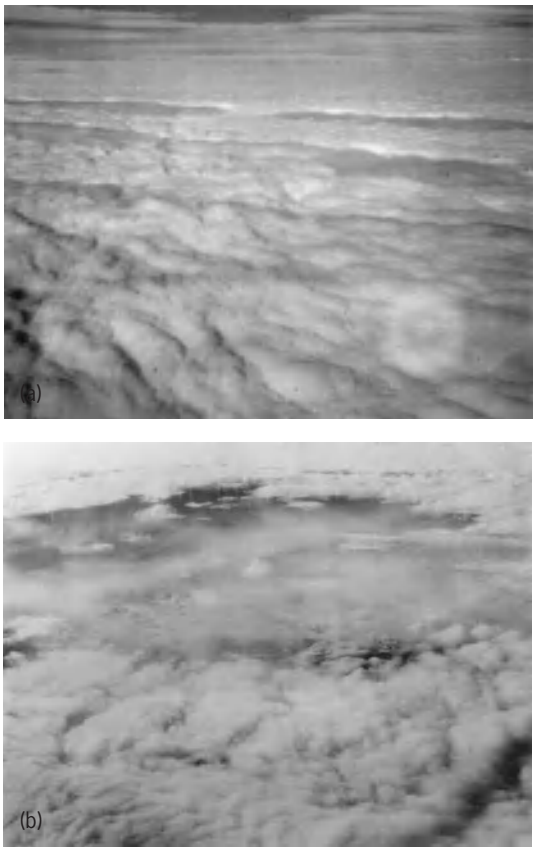


Fig. 3. Cloud dissipation. (*a*) Three lines in a stratocumulus cloud layer 15 min after seeding. (*b*) Opening in the stratocumulus layer 70 min after seeding. (*U.S. Army ECOM, Fort Monmouth, New Jersey*)

It is evident that ice phase modification is possible only when ice crystals are not provided by natural processes—either by direct nucleation or by fallout or downward mixing of ice crystals from colder regions aloft into supercooled clouds below. A lack of crystals is often the case for ground fog or stratus cloud below a clear sky; also, for a field of developing cumuli. Developing clouds resulting from orographic lift over a mountain range offer similar opportunities implemented in many cloud seeding projects undertaken in the Sierra Nevada (California) and the Rocky Mountains. Statistical evaluation of several studies of convective cloud modification has found inconclusive results (Colorado, Florida, Switzerland), and it is evident that the possibilities of seeding for enhanced rainfall are less obvious than had first been thought.

On occasion, natural processes produce secondary ice particles from primary particles resulting from nucleation on mineral aerosol. Growth of soft hail (graupel) from supercooled cloud having an adequate spread of droplet sizes at temperatures

between 24.8° and 17.6°F (−4 and −8°C) gives such particles which rapidly grow to columns some 50–100 micrometers long (Hallett-Mossop process). Under favorable conditions, one ice particle is produced for each 50 m (164 ft) fall of a soft hail particle in the cloud, a sufficient number ot make a significant enhancement of ice concentration and ultimately precipitation. A very small number of particles—well below the detection limit in aircraft measurement—reproduce exponentially under the right conditions. A further process occurs when ice particles fall into dry or warmer air and evaporate or melt, resulting in small ice particles which may be recycled into the cloud in weak updrafts. These considerations suggest that ice cloud seeding can be effective only under a rather limited range of conditions—when natural ice production processes are not effective. Such situations occur when cool bases (<41°F; 5°C) and continental aerosol are present, where ice produced from low temperatures aloft is not recycled.

The pattern of subsequent weather modification attempts has involved searching for situations with regions having large amounts of supercooled water which persist for several hours, seeding in a controlled way, and evaluating the effect produced both by aircraft and by radar. It is being realized that seeding opportunities may be quite narrow. As a result, to become more effective in increasing precipitation may require more skill in locating seeding agents in developing clouds than is currently available. *See* CLOUD PHYSICS; FOG; PRECIPITATION (METEOROLOGY).

Hail modification. Efforts have also been made to reduce hail by similar seeding techniques. This could be accomplished by freezing every droplet in a cloud so that hailstones no longer grow, a technique that would require an unrealistic amount of seeding aerosol and would present serious distribution problems. An alternative and more realistic approach would reduce the sizes of all hailstones so that they are less than a few millimeters in diameter on falling to the ground. This would minimize crop damage. It has been suggested that hail modification can be achieved by seeding the hail-forming part of the cloud so that a given amount of cloud water is distributed over a larger number of smaller stones. Careful control of induced ice concentrations in different parts of the supercooled liquid water cloud would be required, depending on the natural hailstone concentration. This is well beyond available technology. *See* HAIL.

Hurricane modification. It has been suggested that hurricanes might be seeded, either to reduce their maximum wind strength or to change their direction of motion. However, research has revealed that large quantities of ice are often present naturally. Thus systematic large-scale seeding would not be effective. Several hurricanes were seeded during the period 1950–1975 with inconclusive results. Regions of supercooled water cloud do exist asymmetrically around the eye, associated with strong updrafts and sometimes frequent lightning. This offers the possibility of "leading" the hurricane in a different direction by selective seeding in one quadrant. Once lighting has occurred in any system, it implies the existence of an efficient ice-producing process and suggests that further seeding may be less than effective. Again, seeding opportunities may thus be very narrow. *See* HURRICANE; LIGHTNING.

Warm cloud modification. Coalescence precipitation occurs by the collision and coalescence of droplets of diameter greater than 30 μm formed by condensation with slower-falling, smaller droplets. Continuing coalescence eventually results in raindrops several millimeters in diameter. Such larger drops form on occasional large hygroscopic nuclei of sodium chloride (NaCl) or ammonium sulfate [$(NH_4)_2SO_4$] or by special mixing and evaporation mechanisms. This process produces rain not only in clouds whose temperature lies entirely above the freezing point ("warm" clouds), as is often the case in Hawaii, but sometimes also occurs in supercooled clouds. It is evident that many clouds fail to rain at all, and it has been suggested that the larger cloud droplets are not present in sufficient numbers and that seeding with larger droplets might initiate precipitation. Hygroscopic materials—such as sodium chloride—have been dispersed in airborne trails to test such hypotheses; local effects in the cloud microphysics occur, but subsequent enhancement of precipitation remains to be demonstrated. A further hypothesis is that hygroscopic seeding may influence the drop size distribution to provide conditions appropriate for secondary ice formation; while this is possible, the task of modifying the whole of the cloud in just the right place appears formidable. The inverse process—of producing many small droplets by additional nuclei from combustion such as agricultural burning—has also been suggested as leading to a rainfall decrease; this hypothesis is also unproven. Attempts to clear warm fog by high-pressure, high-volume fire-fighting sprays similar in size to 0.08-in. (2-mm) raindrops have shown some measure of success, doubling the visibility over a region of a few hundred yards.

The use of sound—as shock waves from explosives carried aloft on balloons or by ringing church bells—characterized nineteenth-century efforts in rainmaking and hail dissipation. It appears to have had only psychological value, and has been shown to be quite ineffective in causing significant changes in the cloud, either by nucleating ice crystals or by causing droplet coalescence.

Heating effects. Somewhat greater cloudiness and precipitation sometimes exist near cities, which act as "heat islands" and give enhanced convective activity and deeper clouds later in the afternoon. The effect has been found both in Chicago and St. Louis; it results from an increase in the amount of sunlight absorbed by roads and buildings and changes of surface wind flow. More dramatic effects occur over natural extended heat sources such as forest fires and also volcanoes (**Fig. 4**). Under certain conditions, setting a prairie fire will produce a thunderstorm which will extinguish the fire; this skill was practiced only in past times. *See* TERRESTRIAL RADIATION.

Fig. 4. Convective cloud forming over a forest fire.

Radiation modification. Particulates produced by forest fires and by fossil fuel combustion in industrial society pass into the air and are efficiently removed during natural precipitation processes. These particles influence the radiation balance and may be responsible for reducing sunlight and local visibility. The effect may be present on a worldwide basis; particulates have a half-life before removal of about 1 week, and can be carried considerable distances during this time. The overall result is a decrease in direct sunlight, which is somewhat compensated for by an increase in diffuse radiation scattered from other directions. The net result is a decrease in surface radiation available for photosynthesis by about 5%. The atmosphere above is heated by comparable amounts, and calculations show that this heat is lost to space more rapidly than if it were absorbed at the surface; the net result is a slight atmospheric cooling.

Average anthropogenic particle loading is around 20% of natural loading, and each is substantially less than peaks resulting from major volcanic eruptions. It has been suggested that a large effect could result from smoke produced by widespread nuclear war—the nuclear winter. This would lead to cooling the Earth's surface and heating aloft. Particles that enter the stratosphere may remain there several years and influence the radiation over this time. Similar suggestions were made for the effect of the smoke from the Kuwait oil fires following the Gulf War. This smoke was lofted only to about 3 mi (5 km), so it did not enter the stratosphere and was removed by precipitation within a few days along its trajectory; there was no detectable global effect. Contrails from high-level aircraft give a similar effect, increasing high-level cloudiness (optical thickness) in well-traveled parts of the globe. The increasing frequency of travel by jet aircraft in regions of the Northern Hemisphere could ultimately be of global importance. The effect of widespread injection of exhaust nuclei from jet aircraft at high levels (6–7 mi or 10–12 km) may be important for changing the opacity of existing cloud and for serving as particles for chemical reactions. The residence time in the upper troposphere is longer (a few weeks) and may lead to enhanced opacity of cirrus in well-traveled routes.

Greenhouse effect. In addition to particulate emissions, fossil fuel combustion increases atmospheric content of carbon dioxide, about 50% of the emissions remaining in the atmosphere, the remainder being absorbed by the oceans and contributing to enhanced photosynthesis. Carbon dioxide (CO_2) has increased by 15% since 1850, with further increases projected until fossil fuel becomes exhausted in the twenty-first century (**Fig. 5**). Other gases (for example, methane produced by rice paddies) may have a smaller (20%) but similar effect (**Fig. 6**). There is a suggestion that the recent upward tendency in the Earth's temperature may be an indication of this effect (**Fig. 7**). Calculations confirm that increased carbon dioxide would lead to an atmospheric warming, but this would be at least in part offset by the radiational cooling from articles and additional cloud formation. Such calculations are suggestive, but not sufficiently precise to predict the overall effect with confidence.

Aerosol from human activities and from episodal natural sources such as extensive forest fires and large volcanic events may have a more important worldwide effect than originally thought. Larger, more hygroscopic aerosol particles (cloud condensation nuclei) may compete with the natural background produced by ocean spray, gas reactions in the atmosphere from ocean-produced dimethylsulfide, and more universally occurring small forest fires. The result is clouds with a larger concentration of smaller droplets and of greater areal extent.

Urban pollution gives rise to this effect in low-level stratiform clouds as occur off the California coast and in the East Atlantic from industrial pollution from Europe. High-level cirrus clouds are similarly influenced by volcanic aerosols and possibly jet aircraft exhaust. Such clouds result in a higher albedo—more sunlight is scattered back to space. Such clouds form a thicker blanket and result in the emission of less radiation to space because they are colder than the underlying surface. Such changes may significantly reduce the greenhouse effect resulting from increased CO_2. *See* AIR POLLUTION; ALBEDO; GREENHOUSE EFFECT.

Ozone hole. Ozone (O_3) is normally produced at altitudes between 12 and 30 mi (20 and 50 km) in the atmosphere by ultraviolet light absorption by ordinary atmospheric oxygen. The ozone absorbs further ultraviolet radiation at these levels, providing a shield to such radiation, which has harmful biological effects at the Earth's surface.

Measurements of ozone over Antarctica in the late 1980s showed systematic wintertime decreases at 6–12 mi (10–20 km). This has been attributed to the use of manufactured fluorocarbon compounds, which are used in refrigerators, in aerosol sprays, and in the manufacture of foamed insulation. These gases are unreactive in the lower atmosphere; however, at high altitudes they become dissociated and

react with ozone. The rate of reaction is increased by adsorption on particles in the presence of nitrogen oxides (produced worldwide by soils, lightning, fertilizer use, and combustion). This leads to ozone depletion. As ozone loss extends to sunlit regions of the globe, enhanced skin cancer rates and possible damage to plants and algae may result.

The presence of aerosol at high levels in the stratosphere (15 mi or 25 km) where temperatures are low (−139°F or −95°C) leads to the presence of polar stratospheric clouds in the form of nitric acid trihydrate and, at a few degrees colder, ordinary water ice. These particles provide a surface on which chemical reactions leading to ozone destruction occur. Under normal conditions the cloud particles grow to a size of about 50–100 μm and fall out to lower levels and higher temperatures where they evaporate; the clouds lack persistence. In the presence of aerosols of volcanic origin in high concentrations (as after the Pinatubo eruption), reaction rates are much greater because of the larger surface area available; also the more numerous particles are smaller and fall out more slowly. Thus, should low temperatures occur and form clouds, the ozone destruction rate is much greater. The low temperatures are produced by dynamical (lifting) processes and, to a lesser extent, radiation loss in the winter pole; the frequency is greater in the south because the wintertime circumpolar vortex is steadier and more symmetric than in the north, where the upper-level flow is more influenced by the continents. It follows that although the ozone loss in the south is greater, the ozone loss in the north is carried toward midlatitudes because of the enhanced north/south flow in waves around the polar vortex.

The so-called depth of the ozone hole (measured in Dobson units, related to the total amount of ozone) and the area of the effect have been increasing irregularly since the hole was discovered in 1985 above Antarctica. It has subsequently been found above the Arctic. The actual effect, besides depending on the chemical composition of the stratosphere, depends on the amount and area of cooling, on the mixing in and out of the vortex, and somewhat unpredictably, on the aerosol concentration.

Ozone is also produced and reaches damaging concentrations during periods of air pollution in quiescent air as in the Los Angeles Basin in summer or in an extensive slow-moving anticyclone. In this case, ozone results from a sunlight-induced reaction requiring the presence of volatile organic compounds and nitrogen oxides. Important is the presence of a temperature inversion at a height of about 0.6 mi (1 km) to trap the products of the reactions. Urban pollution provides the organic compounds and the nitrogen oxides; over a much wider area, agriculture may also provide nitrogen oxides from fertilized soil and combustion. Such situations exist in the eastern United States and in regions of Europe, China, and Japan. *See* AEROSOL; STRATOSPHERE.

Acid rain. Sulfur is a trace component of the atmosphere, being part of the biological cycle of plants and animals. Emissions of sulfur dioxide (SO_2), particularly from fossil fuel combustion and ore refining and reduction processes, add about half as much again on a worldwide basis. Problems arise from very high concentrations of sulfur dioxide produced in industrial areas such as the northeast United States and the Ruhr Valley in Europe. These gases eventually react to form sulfate ions (SO_4^{2-}), a process enhanced by the presence of nitrogen oxides from automobile exhausts. The reactions take place in cloud

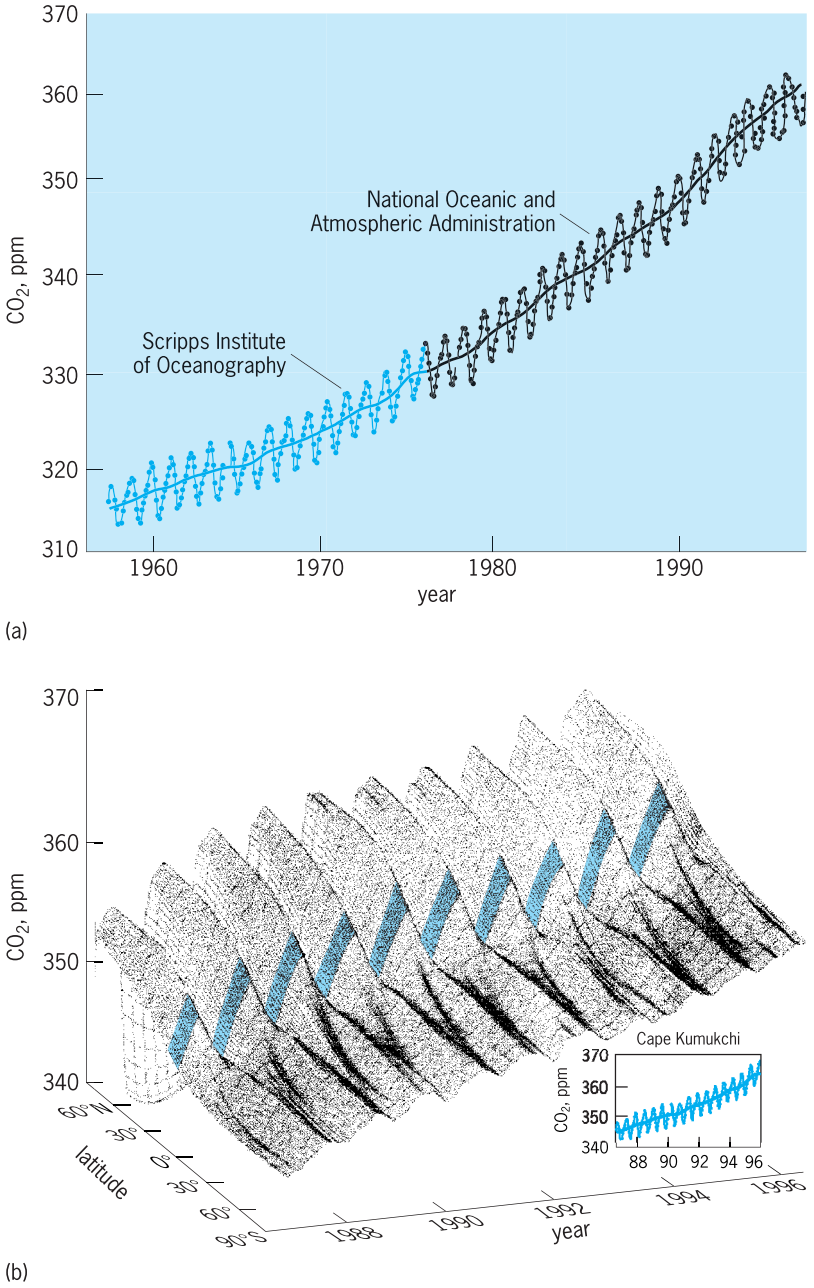


Fig. 5. Carbon dioxide concentrations. (*a*) At Mauna Loa, Hawaii. Annual fluctuations result from seasonal release and absorption of CO_2 by vegetation, primarily in the larger landmasses of the Northern Hemisphere. The steady increase is evident, as are fluctuations of the increase from year to year, probably dependent on climate effects on vegetation behavior. (*b*) Wordwide, from National Oceanic and Atmospheric Administration measurements. Note lower values in the Southern Hemisphere, showing lack of transport on an annual basis of air between the two hemispheres. The annual cycle is still present, but less in magnitude. (*After National Oceanic and Atmospheric Administration; Scripps Institute of Oceanography*)

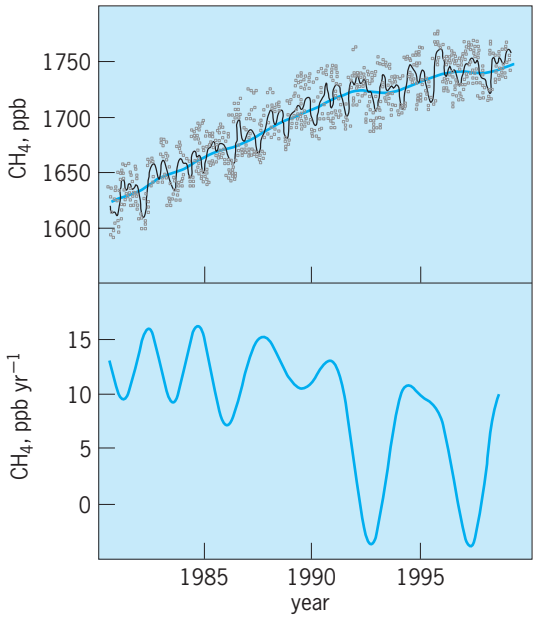


Fig. 6. Concentration of methane CH_4 gas, measured at Mauna Loa Observatory. Note an increasing tendency in the first decade of measurements, lessening in the second decade. The annual cycle is more complex than for carbon dioxide, suggesting a more complex vegetative source. (*After National Oceanic and Atmospheric Administration*)

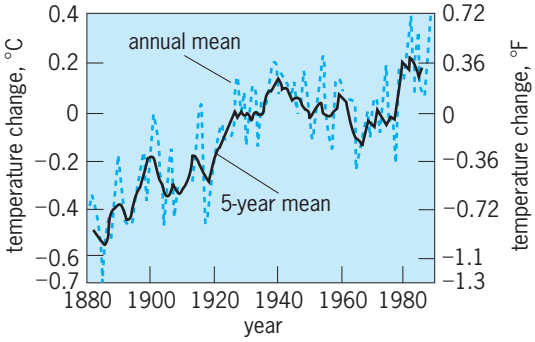


Fig. 7. Global temperature changes, 1880–1987. (*After R. A. Kerr, The weather in the wake of El Niño, Science, 242:883, 1988*)

droplets, which are removed during winter frontal precipitation and summer convective showers, to be deposited as acid precipitation often hundreds of miles downwind of the source. Rain is always slightly acidic (pH 5.7) because of the absorption of atmospheric carbon dioxide. The effect of pollution is to increase this acidity; values in the range pH 2–3 have been measured. It has been suggested that this leads to damage in spacific types of vegetation (some pines and conifers) growing in soils which lack buffering capacity, and to kills of fish in acidic spring runoff. Low-pH fogs have occasionally been observed in the Los Angeles Basin. Freezing on tress of acid fog containing nitrate (NO_3^- and sulfate ions may lead to enhanced acidity and damage at temperatures a few degrees below freezing. The actual extent of the effects is subject to disagreement because of uncertain measurement techniques of earlier years and lack of systematic observations of tree damage from natural causes over a long time period. Further uncertainty results from less frequent (and more intense) forest fires because of control strategies, leading to older stands of trees that are more susceptible to damage. *See* ACID RAIN; PH.

John Hallett

Bibliography. B. Bolin et al., *The Greenhouse Effect, Climatic Change, and Ecosystems*, 1986; A. S. Dennis, *Weather Modification by Cloud Seeding*, International Geophysics Series, vol. 24, 1980; T. Conway et al., Evidence for interannual variability of the carbon cycle from the National Oceanic and Atmospheric Administration/Climate Monitoring and Diagnostics Laboratory Global Air Sampling Network, *J. Geophys. Res.*, 99(D11):22,831–22,855, 1994; J. T. Houghton et al. (eds.), *Climate Change 1995: The Science of Climate Change*, 1995; National Research Council, *Rethinking the Ozone Problem in Urban and Regional Air Pollution*, 1991; S. F. Singer, *The Changing Global Environment*, 1975; C. Spence, *The Rainmakers: American Pluviculture to World War II*, 1980; World Meteorological Organization and United Nations Environmental Program, *Developing Policies for Responding to Climatic Change*, April 1988.

Weather observations

The measuring, recording, and transmitting of data of the variable elements of weather. In the United States the National Weather Service (NWS), a division of the National Oceanic and Atmospheric Administration (NOAA), has as one of its primary responsibilities the acquisition of meteorological information. Most observations are taken by federal employees, although private citizens and private organizations such as the utilities are also involved. The data are sent by various communication methods to National Weather Service field offices and to the National Meteorological Center near Washington, D.C.

At the Center, the raw data are fed into large computers that are programmed to plot, analyze, and process the data and also to make prognostic weather charts. The processed data and the forecast guidance are then distributed by special National Weather Service systems and conventional telecommunications to field offices, other government agencies, and private meteorologists. They in turn prepare forecasts and warnings based on both processed and raw data. *See* WEATHER MAP.

Surface observations. A wide variety of meteorological data are required to satisfy the needs of meteorologists, climatologists, and users in marine activities, forestry, agriculture, aviation, and other fields. This has led to a dual surface-observation program: the Synoptic Weather Program and the Basic Observations Program. *See* AERONAUTICAL METEOROLOGY; AGRICULTURAL METEOROLOGY; INDUSTRIAL METEOROLOGY.

The Synoptic Weather Program is designed to assist in the preparation of forecasts and to provide data for international exchange. Worldwide surface observations are taken at standard times [0000, 0600,

1200, and 1800 Universal Time Coordinated (UTC)] and sent in synoptic code.

The Basic Observations Program routinely provides meteorological data every hour. Special observations are taken at any intervening time to report significant weather events or changes. The vast majority of stations involved in these observing programs in the United States are operated by National Weather Service and Federal Aviation Administration personnel and are referred to as first-order stations. Observation sites are located primarily at airports; a few are in urban centers. At these sites, human observers report the weather elements.

Present weather. This consists of a number of hydrometeors, such as liquid or frozen precipitation, fog, thunderstorms, showers, and tornadoes, and of lithometeors, such as haze, dust, smog, dust devils, and blowing sand. The amount of cloudiness is also reported. *See* FOG; METEOROLOGICAL OPTICS; PRECIPITATION (METEOROLOGY); SMOG; THUNDERSTORM; TORNADO.

Pressure. This measurement is read from either a mercury or precision aneroid barometer located at the station. A microbarograph provides a continuous record of the pressure, from which changes in specific intervals of time are reported. Pressure changes are frequently quite helpful in short-range prediction of weather events. *See* AIR PRESSURE.

Temperature and humidity. These are measured by a hygrothermometer, located near the center of the runway complex at many airport stations. The readings are transmitted to the observation site. Readouts are either digital or dial. The temperature dial indicator is equipped with pointers to determine maximum and minimum temperature extremes; after the extremes are recorded, pointers are normally reset every 6 h. At other sites, temperature and humidity (wet-bulb temperature) are read from mercury thermometers in ventilated wooden shelters located in a grassy area. *See* HUMIDITY; HYGROMETER; TEMPERATURE MEASUREMENT.

Wind speed and direction. These measurements are telemetered into most airport stations. The equipment, consisting of an anemometer and a wind vane, is located near the center of the runway complex at participating airports; elsewhere it is placed in an unsheltered area. *See* WIND MEASUREMENT.

Visibility. This is determined from markers at known distances during daylight or from lights at known distances at night. At jet terminals, transmissometers parallel to the instrumented runway are calibrated in units of runway visibility. At airports equipped with high-intensity runway lights, runway visual range—the distance a pilot can see high-intensity runway lights along the runway while approaching for a landing—is reported. This is determined by a computer, using transmissivity, background illumination, and the intensity setting of the runway lights.

Clouds and ceilings. Various types of clouds and their heights are reported. The lowest height of opaque clouds covering half or more of the sky is known as the ceiling, and is normally measured by a ceilometer at first-order stations. The ceilometer projects a modulated light on the cloud base. At a known distance, the spot of light is electronically scanned day and night, and the cloud height is computed trigonometrically. *See* CLOUD.

Automated surface observing system. The National Weather Service, Federal Aviation Administration, and Department of Defense have been jointly engaged in the development and implementation of the Automated Surface Observing System, which is directed at providing surface observations through the use of modern sensor and computer technology. These systems measure a variety of weather elements, perform quality control, and process, display, and communicate surface information through automated techniques. These agencies have implemented more than 1000 systems in the United States to provide near-continuous (up to every minute) observations of most of the elements that have been provided manually.

Some of the information provided by the Automated Surface Observing System differs from that provided by human weather observers. For example, with this system visibility is measured at one or more visibility sensor locations rather than as the prevailing visibility (representative of a large area) observed by a human. Another difference is in the cloud observations. The human weather observer looks all around the horizon and determines how much of the sky is covered by clouds and how high the cloud bases are—a kind of area integration. The cloud sensor of the Automated Surface Observing System keeps track of clouds that move directly overhead. Its processing system tracks the clouds that have passed overhead over a 30-min period and uses that information to estimate cloud heights and amounts over a larger area—a time integration. Other information from satellites, weather radar, and lightning detection systems supplements observations made by the Automated Surface Observing System.

Other surface measurements. Several of the National Weather Service stations also measure solar radiation and the ozone content of the air, and they collect precipitation samples for chemical analysis. Automated meteorological observing stations have been established to provide basic weather observations at locations too remote for human operation, at stations staffed with part-time personnel, or where complete observation programs are not needed.

In order to better serve the interest of special users, several second-order stations have been established. For example, the U.S. Coast Guard provides limited observations along the coastline every 3 h; local airport operators provide aviation weather observations when needed for aircraft operations; and thousands of cooperative citizens record temperatures and precipitation for climatological purposes. Many of the cooperative stations also report snow density, river stage, and rates of stream discharge and evaporation. Over ocean and lake areas synoptic reports are received from moving ships.

Upper-air observations. Since the 1930s, upper-air observations have been made by the National

Weather Service with radiosondes. The radiosonde is a small, expendable instrument package that is suspended below a 6-ft-diameter (2-m) balloon filled with hydrogen or helium. As the radiosonde is carried aloft, sensors on it measure profiles of pressure, temperature, and relative humidity. These sensors are linked to a battery-powered, 300-mW radio transmitter that sends the sensor measurements to a sensitive ground receiver on a radio frequency ranging 1668.4–1700.0 MHz. By tracking the position of the radiosonde in flight with a radio direction finder or radio navigation system, such as Loran or the Global Positioning System (GPS), information on wind speed and direction aloft is also obtained.

The radiosonde flight can last in excess of 2 h, and the radiosonde can rise over 100,000 ft (30 km) and drift more than 125 mi (200 km) from the release point. The radiosonde is exposed to temperatures as cold as −121°F (−85°C) and air pressures only a few hundredths of that on the Earth's surface.

Although all the data from the flight are used, data from the surface to the 400-millibar (40-kilopascal) pressure level (about 23,000 ft or 7 km) are considered minimally acceptable for National Weather Service operations. Thus, a flight may be deemed a failure and a second radiosonde released, if the balloon bursts before reaching that pressure level or if more than 6 min of pressure or temperature data between the surface and that pressure level are missing.

When the balloon has expanded beyond its elastic limit and bursts (at a diameter of about 20 ft or 6 m), a small parachute slows the descent of the radiosonde, minimizing the danger to lives and property. Only about 15% of the approximately 75,000 radiosondes released by the National Weather Service each year are found and returned to the Service for reconditioning. These rebuilt radiosondes are used again, saving the Service the cost of a new instrument.

Worldwide, there are over 900 upper-air observation stations. Most are located in the Northern Hemisphere, and observations are usually taken at the same time each day (0000 or 1200 UTC), 365 days per year. Observations are made by the National Weather Service at 92 stations—69 in the conterminous United States, 13 in Alaska, 9 in the Pacific, and 1 in Puerto Rico. Data are exchanged between countries through international agreements.

Understanding and accurately predicting changes in the atmosphere requires adequate observations of the upper atmosphere. Radiosonde observations, plus routine aircraft reports, radar, and satellite observations, provide meteorologists with a three-dimensional picture of the atmosphere. *See* LORAN; METEOROLOGICAL INSTRUMENTATION; SATELLITE NAVIGATION SYSTEMS; WEATHER OBSERVATIONS; WIND MEASUREMENT.

Radar observations. National Weather Service, Federal Aviation Administration, and Department of Defense weather radars distributed throughout the United States are used to observe precipitation within a radius of about 250 nmi (460 km), and associated wind fields (utilizing the Doppler principle) within about 125 nmi (230 km). The primary component of this set of weather radars is known as NEXRAD (Next Generation Weather Radar), but significant weather radar observations are also provided by various Federal Aviation Administration systems. All of these radars are operated in an automated, computer-managed manner (with oversight by human personnel) to detect weather targets and provide standardized weather information products. The products provide meteorologists and other users with information on rainfall intensity, likelihood of tornadoes or severe thunderstorms, projected paths of individual storms (both ambient and within-storm wind fields), and heights of storms. This information is most helpful for short-range (up to 3 h) forecasts and warnings. The products are distributed to offices of the National Weather Service, Department of Defense, and Federal Aviation Administration and to a wide variety of other governmental and commercial users, including broadcast media. *See* DOPPLER RADAR; RADAR METEOROLOGY.

Satellite observations. Geostationary weather satellites near 22,000 mi (36,000 km) above the Earth transmit pictures depicting the cloud cover over vast expanses of the hemisphere. Using still photographs and animated images, the meteorologist can determine, among other things, areas of potentially severe weather and the motion of clouds and fog. In addition, the satellite does an outstanding job of tracking hurricanes over the ocean where few other observations are taken. Polar orbiting satellites also provide imagery for forecaster use and temperature soundings to supplement ground-based observations in numerical prediction models of the atmosphere. These data are used not only within the United States but also by many other nations of the world engaged in weather forecasting. *See* HURRICANE; METEOROLOGICAL SATELLITES.

Lightning detection systems. The technology for detecting and locating lightning strikes developed in response to a variety of needs, ranging from safety of space vehicle launches to utility company operations to weather analysis and forecasting. Ground-based lightning detection systems, which cover most of the United States and some other countries, detect the electromagnetic wave that emanates from the lightning path as the lightning strikes the ground. By the year 2000, it is anticipated that satellites will carry optical devices to detect all types of lightning over the Earth. Meanwhile, private organizations and government agencies have been cooperating in implementing lightning detection systems and in applying lightning information to both operations and research. Lightning information has proven to be operationally valuable to a wide variety of users and as a supplement to other observing systems, particularly radar and satellites. *See* LIGHTNING; LIGHTNING AND SURGE PROTECTION.

Density of observation networks. Spacing of observation points is strongly influenced by population density, with many observations near coasts, in valleys, and along main transportation routes. The required density of stations varies for each meteorological user, and there is no universally acceptable

and economically realistic plan. However, meteorologists agree on broad objectives for first-order stations. Roughly, these are surface stations not more than 90 mi (200 km) apart, climatological stations not more than 25 mi (40 km) apart, and upper-air and radar stations not more than 200 mi (320 km) apart. There is a need for a much denser real-time network to complement the first-order stations in order to improve short-range forecasts and warnings. *See* MESOMETEOROLOGY; METEOROLOGY; WEATHER FORECASTING AND PREDICTION.

Frederick S. Zbar; Ronald L. Lavoie

Weathering processes

The response of geologic materials to the environment (physical, chemical, and biological) at or near the Earth's surface. This response typically results in a reduction in size of the weathering materials; some may become as tiny as ions in solution.

Types. The agents and energies that activate weathering processes and the products resulting therefrom have been classified traditionally as physical and chemical in type. In classic physical weathering, rock materials are broken by action of mechanical forces into smaller fragments without change in chemical composition, whereas in chemical weathering the process is characterized by change in chemical composition. In practice, the two processes commonly overlap, almost inseparably. For example, diminution in particle size facilitates chemical reactivity, and an increase in volume of the products during chemical reaction may physically disintegrate the reactants.

Viewed broadly, environments of weathering and the suite of products from each may be categorized in terms of climate, such as desert, arctic, or tropical rain forest. In cold and dry climates, physical weathering predominates and produces angularity in both rock particles and surficial landforms. In warm humid climates, chemical and biochemical weathering yields rounded rock masses, and hydrated and oxidized mineral compounds which may be developed at great depths.

Agents. Within each environment, specific agents of weathering may be recognized and correlated with the types of effects they produce. Important agents of weathering are water in all surface occurrences (rain, soil and ground water, streams, and ocean); the atmosphere (H_2O, O_2, CO_2, wind); temperature (ambient and changing, especially at the freezing point of water); insolation (on large bare surfaces); ice (in soil and glaciers); gravity; plants (bacteria and macroforms); animals (micro and macro, including humans). Human modifications of otherwise geologic weathering that have increased exponentially during recent centuries include construction, tillage, lumbering, use of fire, chemically active industry (fumes, liquid, and solid effluents), and manipulation of geologic water systems.

Products. Products of physical weathering include jointed (horizontal and vertical) rock masses, disintegrated granules, frost-riven soil and surface rock, and rock and soil flows.

Products of chemical weathering include many which have been widely adapted to important economic and technologic uses. Such products include the soil, and the clays used in making ceramic structural products, whitewares, refractories, various fillers and coating of paper, portland cement, absorbents, and vanadium. These are the relatively insoluble products of weathering; characteristically they occur in clays, siltstones, and shales. Sand-size particles resulting from both physical and chemical weathering may accumulate as sandstones.

After precipitation, the relatively soluble products of chemical weathering give rise to products and rocks such as limestone, gypsum, rock salt, silica, and phosphate and potassium compounds useful as fertilizers.

Products of weathering that occur in colloidal sizes, also important qualitatively and quantitatively, are included in the preceding listings.

Processes of chemical weathering. Chemical reactions involving water and gaseous O_2 and CO_2 are probably the most important or abundant weathering processes on Earth. In sharp contrast, on the Moon, which is devoid of such an atmosphere, there is essentially no hydration, oxidation, or carbonation. Aqueous dissolution of rocks and minerals is probably the simplest or most straightforward process of chemical weathering. Solution rapidly removes rock salt (NaCl) and gypsum ($CaSO_4 \cdot 2H_2O$), but more slowly corrodes carbonate, silicate, and oxide rocks.

Hydrolysis. Water dissolves O_2 and CO_2 from the air (possibly 10 times more CO_2 from soil atmosphere), enabling it to oxidize and carbonate, as well as to hydrolyze rocks susceptible to those reactions. For example, Fe in silicate minerals is oxidized to Fe_2O_3, thereby removing Fe from the silicate structure and disrupting that network and making it more vulnerable to further breakdown. Oxidizing water reacts with metallic sulfides to produce the several sulfur-bearing acids, among them sulfuric acid, which is a powerful weathering reagent in itself. The metal constituents of the original sulfides typically become hydroxides or oxides. Fumes containing SO_2, Cl_2, or F_2 from combustion of coal, from smelters, or from industrial furnaces generally combine downwind with water vapor (humidity), rain, fog, or dew to form weathering-effective acids.

Aqueous dissolution of CO_2 produces carbonic acid, which has acidic and complexing (carbonate) properties. Dolostone (dolomite) and limestone (calcite) are quickly dissolved as Ca and Mg bicarbonates in carbonic acid, possibly producing topographic sinkholes, caves, and other karstic features, in addition to erosionally lowering the surface of those rocks. Turbulent and rapid flow of water on carbonate rocks markedly increases the rate of dissolution. The less soluble quartz, chert, clay, or iron oxides contained in dissolving limestone are left behind. Monuments and other structures composed of limestone and marble are similarly attacked.

Silicate rocks are attacked primarily by hydrolysis in a general reaction as shown below, where M refers

$$MAlSiAlO_n + HOH \rightarrow$$
$$M + OH + HSiO_n + H(M)AlSiAlO_n + Al(OH)_3$$

to metal cations (K, Na, Ca, Mg), subscript *n* denotes an unspecified ratio of atoms, and the Al following Si substitutes for $\overline{Si}$. Thus there are formed, by hydrolysis, soluble alkali-metal hydroxides, soluble silica (the ionic distribution depends upon pH), and relatively insoluble clay mineral (or zeolite), or less commonly, hydrated alumina. If the hydrolysis takes place at pH 9.5 or higher, both silica and alumina will be relatively soluble and mobile. They may then be separated and form bauxite ($Al_2O_3 \cdot nH_2O$). Under more acid conditions, clay minerals are formed.

Adding hydrogen ions to the hydrolyzing system increases the rate of reaction. Carbonic acid, formed when the carbon dioxide of the air and soil dissolves in water, is a source of hydrogen ions which accelerate the reaction. Organic (humic) and other acids participate in the hydrolysis. Strongly complexing organic acids may mobilize (complex) in solution Al more effectively than Si from Al-silicate minerals. Solubilization in and precipitation from organic solutions are therefore sensitive to both Eh, the oxidation potential, and pH. Another major source of hydrogen ions is their production in the ionic atmosphere about the rootlets of growing plants. During plant growth and metabolism, hydrogen ions are evolved. These are exchanged by the roots for nutrient cations (K^+, Ca^{2+}, Mg^{2+}) present in nearby clay colloids and rocks. Thus, the process of nutrition of plants is simultaneously a process of weathering of rocks. Hence, the energy which drives plant growth and is indirectly derived from the Sun likewise furnishes some of the energy for weathering of rocks.

Plant activity. Plants that are primitive in development apparently possess higher energies of cation exchange than do those that are more advanced. Lichens derive nutrient cations from fresh rock without intermediary soil. It is difficult to assess quantitatively the extent to which bacteria in the soil, and those coating interstices among mineral grains, accomplish chemical rock weathering, but some pedologists consider bacteria to be a major agent. *See* SOIL MICROBIOLOGY.

Rootlets of macroplants may sorb nutrients from adjacent soil when the mean free-bonding energy of the rootlet exceeds the crystal-bonding energy of mean free-bonding energies of clay minerals or organic substances by which they hold individual nutrient ions in polyionic systems in the soil. Hence, plant nutrition and the activity of agriculture occupy an intermediate position in the weathering sequence between fresh rock-forming minerals and intensely weathered "final" products of weathering (**Figs. 1** and **2**). Chelating organic substances extract cations from rocks, implementing rock breakdown. Partial weathering makes the rock constituents more available to plants, but extended weathering removes the nutrient materials entirely.

Results of chemical weathering. As shown by the hydrolysis reaction, the products from it may be broadly grouped into relatively soluble and relatively insoluble categories. The ultimate destination of the soluble products is the ocean, where they are concentrated in solution or removed by precipitation. Potassium released in solution by weathering, although as soluble as sodium, is more tightly sorbed by clay minerals and may be fixed in crystals of hydrous mica. Dissolved potassium is therefore less abundant than sodium in seawater. Magnesium may

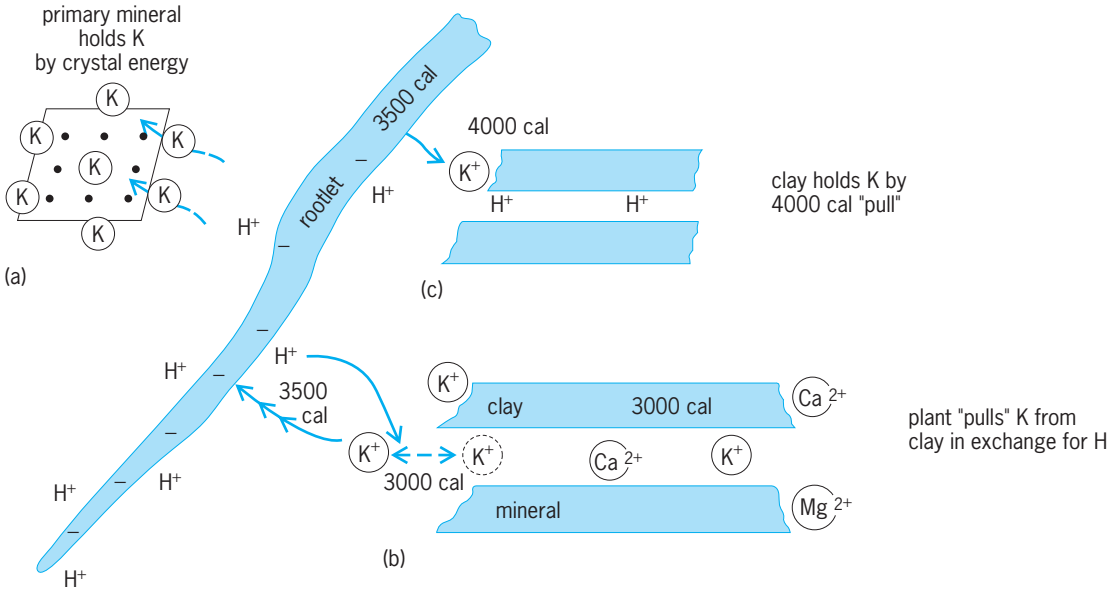


Fig. 1. Exchange-energy relationships between a rootlet and three minerals. (*a*) A potassium-bearing, primary silicate mineral. (*b*) A clay mineral well stocked with exchangeable metal cations. (*c*) A clay mineral scantily stocked with metal cations. The exchange bonding energy (calories per gram-equivalent weight) of K for H in the rootlet exceeds that in only the well-stocked clay mineral which is thus the only one of the three minerals from which nutrient ions can be taken. 1 cal = 4.18 J. (*After W. D. Keller, Mineral and chemical alluviation in a unique pedological example, J. Sediment. Petrol., 31:80–86, 1961*)

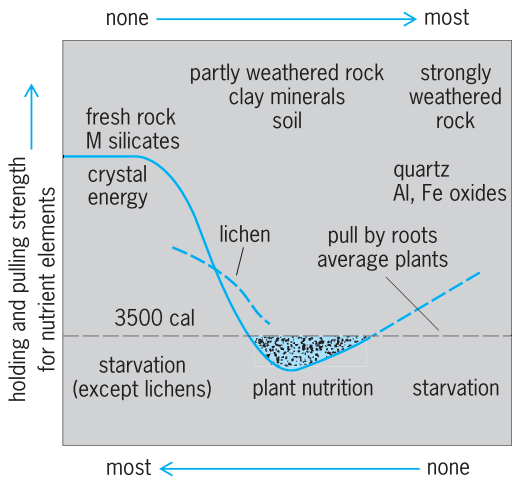


Fig. 2. Comparison of binding energy (calories per gram-equivalent weight) on nutrient metal cation (indicated by M) by rocks and soil minerals with exchange binding energy on cation by plant roots. The relation of plant nutrition to weathering and abundance of nutrients is shown. 1 cal = 4.18 J. (*After W. D. Keller, Mineral and chemical alluviation in a unique pedological example, J. Sediment. Petrol., 31:80–86, 1961*)

be incorporated in chloritic varieties of clay minerals. *See* CLAY MINERALS.

The most abundant weathering products of silicate rocks are the clay minerals. Weathering (hydrolysis) taking place in an environment such that high concentrations of calcium, magnesium, and iron (particularly ferrous) are built up tends to produce the smectite group of clays. Such a high concentration of ions occurs where evaporation exceeds precipitation, ground-water drainage is poor, or hydrolysis is rapid (as in weathering of volcanic dust). The kaolin group of clay minerals is developed where rainfall exceeds evaporation and leaching is intense. Oxidation of iron is then ordinarily high. Under conditions of very drastic leaching and continual wetting of the rocks, as in a tropical rain forest, silica and most cations dissolve, leaving hydrated oxides of alumina and ferric iron (bauxite and laterite). Rising ground–water solutions may carry Al and Fe upward and, because of evaporation or oxidation of organic complexes, leave deposits of both in the tropical subsoil. A high K^+/H^+ ratio in the aqueous-weathering system of Al-silicates yields the illite clay mineral (disordered K-mica). Weathering processes apparently reach a state of near-equilibrium with respect to kaolinite or smectite in environments such as those that prevailed where thick, valuable deposits of the clays were formed. In contrast, surface-exposed weathering of boulders and outcrops yields highly varied and changing products, quasimineral compounds, and rock wreckage.

Clay minerals, although relatively stable products of weathering in one environment, may be decomposed if subjected to more drastic leaching in another environment by processes of the removal of exchangeable cations, the more tightly fixed potassium of illite (hydrous mica) and possibly silica. Clay minerals are said to be degraded when their structures are partly destroyed. Entirely desilicated clays become bauxite or laterite. *See* BAUXITE; LATERITE.

Walter D. Keller

Bibliography. D. Atkinson, *Access to Geography: Weathering, Slopes and Landforms*, 2005; W. J. Bland and D. Rolls, *Weathering: An Introduction to the Basic Principles*, 1998; M. J. Johnsson and A. Basu (eds.), *Processes Controlling the Composition of Clastic Sediments*, 1994; A. Lerman and M. Meybeck (eds.), *Physical and Chemical Weathering in Geochemical Cycles*, 1988; R. Littke, *Deposition, Diagenesis, and Weathering of Organic Matter-Rich Sediments*, 1993.

Wedge

A piece of rigid material whose two major surfaces make an acute angle θ. It is closely related to the inclined plane and is used to multiply the applied force and to change the direction in which it acts.

For the simplest case, in which there is no friction, the forces are those shown in **Fig. 1***a*. Force F is the smaller (applied) force and Q is the larger (useful), force to be exerted. In the absence of friction, forces must act normal to the surfaces; thus, the actual force on the inclined surface is not Q but a larger force F_n (Fig. 1*b*). Summing up forces in the vertical and horizontal directions gives Eqs. (1); combining the

$$\begin{aligned} Q - F_n \cos\theta &= 0 \\ F_n \sin\theta - F &= 0 \end{aligned} \tag{1}$$

expressions thus given for F and Q, and solving for

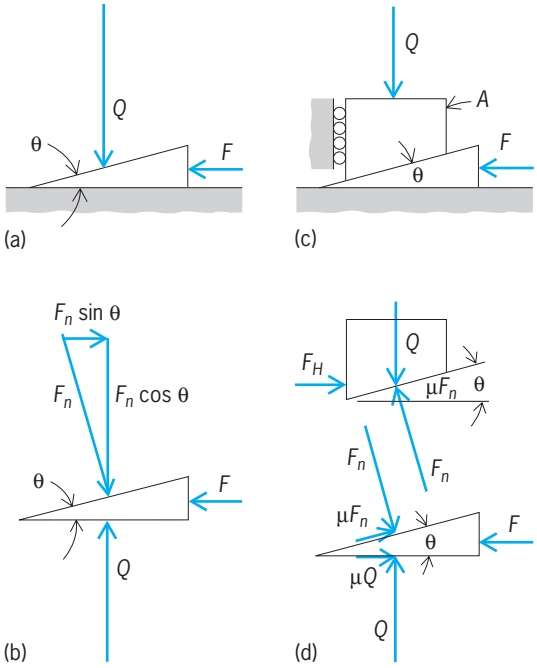


Fig. 1. Forces acting on a wedge : (*a*,*b*) without friction and (*c*,*d*) with friction.

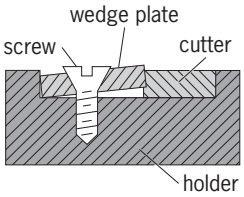


Fig. 2. Wedge plate used as clamp to retain cutter in holder.

F gives Eq. (2). Equation (3) gives the mechanical

$$F = Q \tan \theta \tag{2}$$

$$\text{M.A.} = \frac{Q}{F} = \frac{1}{\tan \theta} \tag{3}$$

advantage (M.A.), the ratio of the useful force to the applied force, as derived from Eq. (2).

In reality, friction is always present, and the analysis becomes considerably more complicated because friction forces will act to oppose relative motion (sliding) between the contacting surfaces. The simplest case is shown in Fig. 1*c*. Here the coefficient of friction μ between the wedge and adjacent members is constant, and the massless block A, upon which Q acts, is constrained without friction to move vertically. The analysis of the forces acting on the block and the wedge, shown in Fig. 1*d*, yields Eq. (4), and the mechanical advantage in this case is given by Eq. (5).

$$F = Q \frac{(1 - \mu^2) \tan \theta + 2\mu}{1 - \mu \tan \theta} \tag{4}$$

$$\text{M.A.} = \frac{1 - \mu \tan \theta}{(1 - \mu^2) \tan \theta + 2\mu} \tag{5}$$

In Eq. (4) it can be seen that the required force increases rapidly as μ increases, becoming infinite when $\mu \tan \theta = 1$, or $\mu = 1/\tan \theta$. At this point the mechanical advantage goes to zero.

When the block A is moving downward, the friction forces, μF_n and μQ, reverse their directions, and force F acts to restrain the motion. For this case, the force is given by Eq. (6).

$$F = Q \frac{(1 - \mu^2) \tan \theta - 2\mu}{1 + \mu \tan \theta} \tag{6}$$

In Eq. (6) it can be seen that the numerator decreases as the coefficient of friction μ increases, and the sign of F can become negative for a sufficiently large value of μ. When F becomes negative, the wedge must be pulled out to permit block A to move downward. Such a system is called self-locking. No simple equation exists for calculating the value of μ required for self-locking, but for practical purposes the system will be self-locking when $\mu \geq (\tan \theta)/2$.

Some applications of the wedge are in splitting wood, in raising the platform of low-lift platform trucks, in cone clutches, in V-belt drives, and in fastening parts together (**Fig. 2**). *See* INCLINED PLANE; MECHANICAL ADVANTAGE; SIMPLE MACHINE.

Richard M. Phelan

Weeds

Unwanted plants or plants whose negative values outweigh the positive values in a given situation. Weeds cost growers millions of dollars each year in reduced yield and quality of agricultural products. Especially in tropical areas, irrigation systems have become unusable because of clogging with weeds. Weeds can harbor deleterious disease organisms and insects that harm crops cost livestock producers millions of dollars. In addition, weeds can cause allergic reactions and serious skin problems (poison ivy), break up pavement, slow or stop water flow in municipal water supplies, interfere with power lines, cause fire hazards around buildings and along railroad tracks, and produce poisonous plant parts. *See* ALLERGY.

Characteristics. The most serious weeds are those that succeed in invading new areas and surviving at the expense of other plants by monopolizing light, nutrients, and water, or by releasing chemicals detrimental to the growth of surrounding vegetation (allelopathy). In the plant kingdom, dozens of species have been shown to release allelopathic chemicals from roots, leaves, and stems. The chemicals can be of many different types, including alkaloids, phenolics, terpenoids, and nitriles. Prolific seed production, seed dormancy, rapid seed germination, vigorous seedling growth, and a rapid, spreading growth habit are all beneficial to a weed for invasion and survival. *See* ALLELOPATHY.

Classification. Weeds can be classified as summer annuals, which germinate in the spring, set seed, and die in the fall (crabgrass); winter annuals, which germinate in the fall, set seed, and die in the spring (common chickweed); biennials, which germinate one year, overwinter, set seed, and die the following summer (wild carrot); simple perennials, which live for several years but spread only by seed (dandelion); and creeping perennials, which live for several years and can spread both by seed and by underground roots or rhizomes (field bindweed).

Control methods. Hand pulling, fire, flooding, and tillage are useful for controlling weeds in many situations. Insects and pathogens have also been introduced to control certain weed species. Techniques such as herbicides, computerization of spray and tillage technology, remote sensing for weed mapping and identification, and laser treatment have also been explored.

Various chemicals have always been used to suppress vegetation, but the discovery of the herbicidal properties of 2,4-D during World War II provided the stimulus for a search for new herbicides, resulting in a steady flow of new chemical tools. There are approximately 150–180 chemicals in hundreds of different formulations sold commercially as herbicides throughout the world. Some are mixed into the soil before the crop is planted, some are applied to the soil surface, and some are applied to the foliage of emerged weeds. Often, weeds can be killed without injury to the crop, because the crop fails to absorb the herbicide through a protective cuticle

on the leaves or a deep root system, degrades the herbicide before it can exert a phytotoxic effect, or contains a different form of a crucial enzyme that the herbicide blocks in the weed.

Herbicides can be applied safely and have resulted in large improvements in the availability and quality of food. They have increased the feasibility of no-till agriculture, leading to significant reductions in soil erosion. They commonly kill weeds by disrupting a physiological process that is not present in animals. Examples are the triazines, ureas, and uracils that block photosynthesis, or glyphosate and the sulfonylureas that block synthesis of essential amino acids. Consequently, many herbicides are very low in toxicity to animals. Some, however, are moderately high in toxicity and must be used carefully. *See* HERBICIDE; PHOTOSYNTHESIS.

Herbicide resistance. Although highly efficient herbicides have ben introduced, there has been a tendency to use the same herbicides continuously over several seasons because they performed more satisfactorily than the previous herbicide. However, rare individual weeds that are genetically resistant to the herbicide have flourished and reproduced, leading to populations of weeds resistant to that herbicide. For example, several dozen species have been identified as resistant to triazine herbicides, and resistant populations have developed in response to sulfonylureas, dinitrobenzeneamines, aryloxyphenoxypropionic acids, and others. Arnold P. Appleby

Biological control. Biological control involves the use of natural enemies (parasites, pathogens, and predators) to control pest populations. In the case of weed pests, the primary natural enemy groups utilized are arthropods, fungal pathogens, and vertebrates. The two major approaches include classical and inundative biological control. Biocontrol can be a highly effective and cost-efficient means of controlling weeds without the use of chemical herbicides.

Classical methods. Classical biocontrol (also termed the inoculative or importation method) is based on the principle of population regulation by natural enemies. Most naturalized weeds leave behind their natural enemies when they colonize new areas, and then can increase to significant densities. Classical biocontrol involves the importation of natural enemies, usually from the area of origin of the weed (and preferably from a part of its native range that is a good climatic match with the intended control area), and their field release as biocontrol agents. Imported biocontrol agents must be host-specific to the target weed.

Inundative methods. Inundative control is the mass production and periodic release of large numbers of biocontrol agents to achieve controlling densities. It can be used where existing populations of agents are lacking or where existing populations that are not self-sustaining at high, controlling densities can be augmented. This method is most frequently employed with fungal pathogens, where typically an agent can be applied as a mycoherbicide. One of the chief advantages of this method is that it can be integrated with conventional farming practices on cultivated croplands.

Arthropod agents. Arthropods, especially insects, are heavily utilized as imported biocontrol agents in uncultivated environments such as grasslands and aquatic systems. The most important insect orders for this are Coleoptera, Lepidoptera, Diptera, and Hemiptera. Among the many weeds successfully controlled by imported insects are prickly-pear cacti (*Opuntia* spp.) in Australia, India and South Africa; Klamath weed (*Hypericum perforatum*), which was reduced to 1% or less of its former abundance in Canada and the United States; tansy ragwort (*Senecio jacobaea*) and alligatorweed (*Alternanthera philoxeroides*) in the United States; and salvinia (*Salvinia molesta*) in Australia. *See* ARTHROPODA; COLEOPTERA; DIPTERA; HEMIPTERA; LEPIDOPTERA.

Fungal agents. Rusts (Uredinales) are the fungal group most frequently employed as imported agents in classical programs. The rust *Puccinia chondrillina*, imported from Italy, controlled the narrow-leaf form of skeletonweed (*Chondrilla juncea*) in Australia. Formulations of spores of foreign and endemic pathogens can be used as mycoherbicides in inundative applications. *See* FUNGI.

Vertebrate agents. Vertebrate animal agents typically do not possess a high degree of host-plant specificity, and their feeding has to be carefully managed to focus it on the target weeds. Goats are sometimes used to control brush in and surrounding urban areas. Fish, especially the grass carp (*Ctenopharyngodon idella*), are used to control submerged aquatic weeds. Sterile, triploid grass carp have reduced hydrilla (*Hydrilla verticillata*) biomass by more than 90% in irrigation canals in southern California. *See* AGRICULTURAL SOIL AND CROP PRACTICES. Charles E. Turner

Bibliography. F. M. Ashton and T. J. Monaco, *Weed Science: Principles and Practices*, 4th ed., 2002; K. L. S. Harley and I. W. Forno, *Biological Control of Weeds: A Handbook for Practitioners and Students*, 1992; H. M. LeBaron and J. Gressel (eds.), *Herbicide Resistance in Plants*, 1982; S. R. Radosevich and J. S. Holt, *Weed Ecology*, 2d ed., 1997; E. C. Rice, *Biological Control of Weeds and Plant Diseases; Advances in Allelopathy*, 1995; D. O. TeBeest (ed.), *Microbial Control of Weeds*, 1991.

Weight

The gravitational weight of a body is the force with which the Earth attracts the body. By extension, the term is also used for the attraction of the Sun or a planet on a nearby body. This force is proportional to the body's mass and depends on the location. Because the distance from the surface to the center of the Earth decreases at higher latitudes, and because the centrifugal force of the Earth's rotation is greatest at the Equator, the observed weight of a body is smallest at the Equator and largest at the poles. The difference is sizable, about 1 part in 300. At a given

location, the weight of a body is highest at the surface of the Earth; it diminishes with altitude and with the depth below the surface. For example, the weight of a body diminishes by about 0.1% if it is raised 2 mi (3 km) above the Earth's surface or taken 4 mi (6 km) below the surface. Weight also depends to a smaller but measurable degree on the density of the Earth's crust below the body. Weight is measured by several procedures. *See* BALANCE.

Since weight is a force, it is expressed in force units. In the United States, the commonest unit of weight is the pound, sometimes written pound force or pound weight, to distinguish it from the mass unit, pound. Pound weight is the weight of a 1-pound mass at a location where the acceleration of gravity is 32.174 ft/s^2 (9.80665 m/s^2). Where the acceleration of gravity is g, the weight of a 1-pound mass is g/32.174 pound weight.

In terms of the international kilogram the pound avoirdupois is now defined as being exactly equivalent to 0.45359237 kilogram. In relation to three smaller avoirdupois weight units, 1 pound equals 16 ounces, 256 drams, and 7000 grains. Since the relation of the pound to the kilogram is divisible by 7, 1 grain equals 0.06479891 gram, exactly.

Besides the avoirdupois pound there is the troy or apothecary pound, in which there are 5760 grains, so that this pound is equal to 576/700 avoirdupois pound. In general, the term pound is taken to mean the avoirdupois pound unless definitely stated otherwise.

The carat is a unit of weight used in evaluating precious stones, and 1 carat equals 200 milligrams. *See* MEASURE. Howard S. Bean

Bibliography. A. V. Astin, *Refinement of Values for the Yard and Pound*, Fed. Regist. Doc. 59-5442, July 1, 1959; D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics*, 6th ed., 2001; L. V. Judson and L. E. Barbrow, *Units and Systems of Weights and Measures: Their Origin, Development and Present Status*, Nat. Inst. Stand. Tech. Lett. Circ. LC 1035, 1985; U. S. Department of Commerce, National Institute of Standards and Technology, *Factors for High-Precision Conversion*, Lett. Circ. LC 1071, 1976.

Weightlessness

A condition induced by the effective lack of resistance to gravitational force on an object or organism, sometimes known as free fall.

Explanation. Isaac Newton provided mathematical explanation for weightlessness in 1687. He was the first person to define gravitation as a force of attraction existing between all bodies, and as such one of the many forces of nature. In his classical theory of gravitation, Newton implied that each body in the universe, from a planet to a molecule, attracts every other body. This gravitational force dictates the orbit of the planets about the Sun, the ocean tides on Earth, and the adherence of humans to Earth. Gravitation is not altered by changes in chemical composition or physical state, as are electrical and magnetic forces; and furthermore, no means have been found to negate gravitational attraction. Newton proposed the law of universal gravitation, which states that two bodies of matter in the universe attract each other with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between their centers. In symbols, this is expressed as $F = Gm_1m_2/d^2$, where F is the gravitational force, m_1 and m_2 the masses of the two bodies, d the distance between them, and G a constant of proportionality called the universal constant of gravity. At the Earth's surface, $F = m_1(Gm_2/d^2) = m_1a$, where m_2 is the Earth's mass, d is the Earth's radius, and F the weight of mass m_1. The average acceleration of gravity $a = Gm_2/d^2$ is approximately 32.2 ft/s^2 or 9.8 m/s^2 at the Earth's surface, which is defined as 1 g. According to this law, even a small increase in the distance between bodies will produce a large decrease in the gravitational force, since the force decreases with the square of the distance. As a body moves from the Earth's surface to a location an infinite distance from the Earth, the gravitational force approaches zero and the body approaches weightlessness. In the true sense, a body can be weightless only when it is an infinite distance from all other objects.

Newton's second law of motion, which states that a force F acting on a body produces acceleration, explains the relationship of gravity to weight. This force acting on a body is equal to the mass m of the body times its accleration a, or $F = ma$. When this law is applied to explain weightlessness, it is evident that the weight force, or the force of gravity, produces the acceleration of gravity, which is the acceleration experienced by a freely falling body. Near the Earth's surface, freely falling bodies fall with a constant acceleration of gravity g, referred to as a standard gravity. It can then be stated that weight W is equal to the mass of the body times the acceleration of gravity, or $W = mg$. Since the weights of different bodies are proportional to their masses, in a uniform gravitational field, weight is frequently but inaccurately used to signify mass. One pound of weight is the force needed at a given location to accelerate a 1-lb mass at a rate equal to the acceleration of gravity at that location. Although weight and mass are both referred to as pounds in the English system of units, they are not the same; mass is the amount of matter present in a body irrespective of its location. Conversely, weight varies with the location of the body with regard to the gravitational field in which it is located (for example, Sun or Earth). Therefore, while the mass of a body stays constant anywhere in the universe, the weight of a body changes as the gravitational force imposed upon it is changed; when that gravitational force is reduced to zero, the body becomes weightless.

Weightlessness is also defined as a condition in which no acceleration, whether of gravity or any other force, can be detected by an object or organism within the system in question. According to Albert Einstein's principle of equivalence, there is no way to distinguish between the force of gravitational

fields and the forces due to inertial motion. When a gravitational force on a body is opposed by an equal and opposite inertial force, a weightless state is produced. This is based on the fact that the mass that determines the gravitational force of a body is the same as the mass related to the acceleration produced by an inertial force of any kind. These inertial force have no external physical origin, but are the consequences of an accelerated state of motion. Because of inertia, a moving object always tends to follow a straight line. When a person swings a bucket by the handle in a large circle, the person feels a pull on the hand, because inertial force (also called centrifugal force in this case) tends to keep the bucket moving in a straight line, while the bucket holder exerts a counterforce constraining the bucket to move along the circle. A similar situation exists in a spaceship orbiting the Earth 200 mi (320 km) above the Earth's surface, where the gravitational field is only slightly weaker than at sea level. The ship is in a state of free fall and therefore weightless. However, while it is falling or in effect being pulled toward Earth by the Earth's gravitational attractive force, the inertial force of the moving ship is directed radially outward from the Earth. Consequently, the ship follows an orbital path. *See* GRAVITATION; GRAVITY; INERTIA; NEWTON'S LAWS OF MOTION.

Realization. Weightlessness is not easily achieved. On Earth, weightlessness can be effectively created for only a few seconds by free fall using a drop tower, or for 20 or 30 s by flying a keplerian parabola in an aircraft. Rockets have been used and improved since the 1950s to achieve low-gravity conditions for scientific studies. Significant biological and materials-science research is conducted in space and on rockets that maintain microgravity conditions for 12–18 min. However, most biological and materials processes occur over a longer time frame, and require space flight to achieve effective weightlessness for periods long enough to conduct meaningful experiments. Both crewless satellites and crewed spacecraft that orbit the Earth for days and even months provide weightless conditions that can truly be used as a tool for basic and applied research (**Figs. 1** and **2**). *See* ROCKET ASTRONOMY; SCIENTIFIC AND APPLICATIONS SATELLITES.

Effects. The physical properties of the space flight environment, predominantly weightlessness, greatly impact the performance of experiments and provide the opportunity to study the effects of gravity. Efforts in fluid physics, combustion research, fundamental physics, material sciences, and biology take advantage of the weightlessness of space flight to better understand processes normally obscured by the presence of gravity.

Biological effects. Gravity is a constantly present major feature of the physical environment on the surface of the Earth. The Earth's gravity exerts a pulling force on all objects, tending to shape or deform them. Galileo recognized that as objects on the Earth's surface increased in mass the gravitational loading or weight increased, which acted to limit the size of the object. Comparing the weight-bearing musculoskeletal systems of different species of land mammals provides an example of gravity acting as a determinant of biologic function and morphology. When a mouse, human, and elephant, with respective masses of 0.7 oz (20 g), 150 lb (70 kg), and 15,000 lb (7000 kg), are compared, the skeletal fractions of their total body mass are seen to increase from 5% for the mouse to 14% for the human and 27% for the elephant; thus, the larger the animal, the larger the supporting-tissue fraction of the total body. Since gravity has persisted throughout the evolution and development of life on Earth, it seems reasonable to speculate that it is intimately involved in the form and function of most, if not all, of the organisms found on Earth. These, as well as other observations, have led to the belief that exposure of organisms to weightlessness would produce profound biological effects.

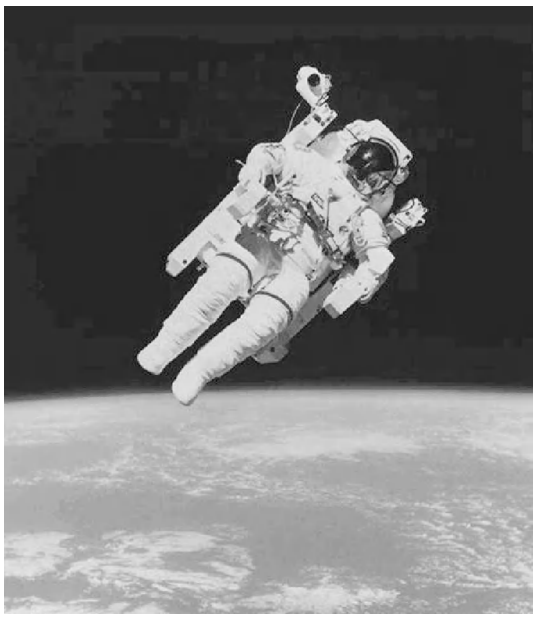


Fig. 1. Weightless astronaut Bruce McCandless during the first free (untethered) extravehicular activity outside the space shuttle *Challenger* in February 1984. Wearing the manned maneuvering unit (MMU) backpack system, McCandless traveled up to 300 ft (90 m) from the orbiter. (*NASA*)

Fig. 2. Astronaut blood flow being monitored on the middeck of the Earth-orbiting space shuttle *Atlantis* flown in October 1989. (*NASA*)

To date, biological experimentation in space and data derived from space have been limited by the number and duration of space flights. Space-flight crews have experienced weightlessness for periods up to a year, and the main body of physiological data on the effects of weightlessness on humans has been derived from observations made on these few crew members. Physiological changes occurred in most systems of the body. The mechanisms involved in most of the observed changes are, however, still unknown.

Experiments using cells have defined a pattern of change in a number of genes. These genes are specific to the structure of the supporting systems within individual cells. There appears to be a down regulation of the genes associated with the cytoskeleton, which internally supports cells. The fact that a body that is the size of a cell possess the means to sense a change in the gravitational environment suggests an amplification of physical forces via an as yet unidentified signaling pathway. Though the response to altered gravity appears to be sensed, other factors, possibly other physical forces, appear to be playing a role in the responses. In muscle tissue, organoids that have a physical load (stretch) placed on them still exhibit changes associated with exposure to the weightless environment. Thus, the changes at the cellular level are due not solely to the absence of gravity (loading) but to other physical properties of the environment. The mechanisms involved in most of the observed changes in biological systems are still unknown. These processes will be investigated on the International Space Station in a number of uniquely designed pieces of equipment such as the European Space Agency's modular culture system. *See* AEROSPACE MEDICINE; SPACE BIOLOGY.

Behavior of materials. In space, weightlessness produces flight problems. In the absence of gravity, the fuels, battery electrolytes, and other liquids do not sink to fixed levels, but coalesce in response to surface tension and the geometry of their containment. Small parts and all installation procedures must be kept scrupulously clean to keep foreign bodies of metallic dust or shavings from floating free during weightlessness and contaminating the atmosphere. Engineering problems exist in the design of feeding and waste management systems as well as laboratory equipment, plant growth chambers, animal holding facilities, and air revitalization and water reclamation systems. All of these must be designed to function in an environment where fluids float in droplets, and unattached or untethered articles drift away. *See* SPACECRAFT PROPULSION; SURFACE TENSION.

In space, sedimentation, buoyancy, and convection do not occur. This allows surface tension to play a primary role in the shape of liquids, and gas and liquid distribution and flow. Further, the role of diffusion and convection in mass transport can be distinguished in a weightless environment. *See* BUOYANCY; CONVECTION (HEAT); DIFFUSION; SEDIMENTATION (INDUSTRY).

The effect of weightlessness on materials offers a revolutionary new opportunity in materials processing. Gravity affects the mixing, separating, welding, heating, and cooling of materials on Earth. In a mixture of molecules, the light ones float to the top while the heavy ones settle to the bottom of the mixture. When light and heavy metals are heated and melted on Earth, the heavier components settle before the mixture can cool to a solid state. In the gravity environment of Earth, convective mixing occurs when heat is introduced into a material during a purifying or refining process. In electrophoresis, a process used to separate, characterize, and analyze certain biological materials, convective currents arise in the presence of gravity, and unwanted particle mixing occurs. Similarly, some characteristics of inorganic materials such as crystalline perfection, homogeneity of precipitation in multiphase systems, and purity are affected adversely by gravity when processing occurs on Earth. *See* ELECTROPHORESIS.

The unusual effect of weightlessness on the behavior of materials was demonstrated on the United States *Skylab* spacecraft in the early 1970s. It was found that in this near-weightless condition, liquid masses formed free-flying globules that could be tethered with a string; vapor grew into unlimited-size bubbles in boiling liquids; combustion products built up around burning materials and smothered the flames (**Fig. 3**); a dispersion of two immiscible liquids was more than 10^5 times more stable in space than on Earth; and reduced gravitational stresses of intramolecular bonding produced unusually large, nearly structurally perfect crystals.

Large germanium selenide crystals were grown on *Skylab*. These space-grown crystals had an almost perfect structure and quality, and one was six times larger than expected. Indium antimonide crystals that were formed in space by melting and then cooling demonstrated the shaping effects of surface tension and adhesion. In a third experiment the prediction that liquid float zones, the liquid zones

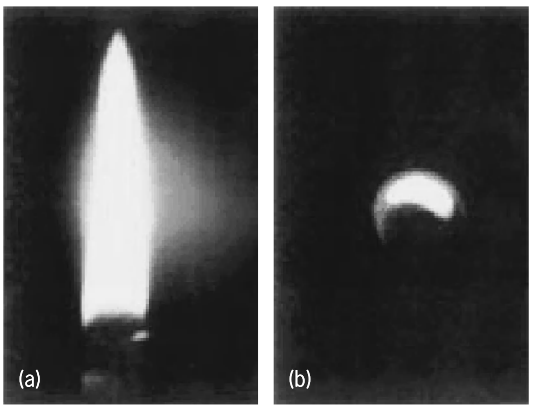


Fig. 3. Effect of weightlessness on combustion. (a) Flame on Earth is elongated. (b) Flame in space forms a circle because the lack of convection (heat rising) does not allow waste material to flow away from the flame. In an environment with no airflow, this causes the candle to burn itself out. (*NASA*)

suspended between two stationary disks, would be longer in space was confirmed. This experiment demonstrated the advantage of conducting floatzone refining, an important process in the manufacturing of high-grade electronics materials, in space. *See* CRYSTAL GROWTH; ZONE REFINING.

A series of protein crystal growth experiments were conducted on the space shuttle starting in 1983. The intent was to eliminate the density-driven convective flow and sedimentation that occurs on Earth and interferes with the formation of single large crystals. On Earth, most biological macromolecules are poorly ordered, diffract poorly, and are not suitable for the high-resolution diffraction analysis that is used to determine their three-dimensional molecular structure. Porcine elastase, interferon, and isocitrate lyase crystals grown in the space shuttle were larger, were more uniform, and yielded diffraction data of significantly higher resolution than the best crystals grown on Earth. This advance will allow scientists to identify the biological function of these macromolecules, which is determined by their structure (**Fig. 4**). *See* CRYSTALLOGRAPHY.

Russia (formerly the Soviet Union) too is actively pursuing materials science in space. A device for smelting bismuth, tin, lead, and cadmium, and a crystal growth experiment were evaluated on *Salyut 5* (1976–1977). On *Salyut 6*, in orbit from 1977 to 1982, several materials-processing furnaces were appraised and infrared-sensitive semiconductors were produced. Superconductors, eutectics, alloys, pure metals, glass, ionic crystals, and metal oxides were produced in other experiments. The first long-duration crew of *Salyut 7*, launched in April 1982, used a new computer-controlled 300-lb (135-kg) materials-processing furnace to produce several pounds of semiconductor monocrystals.

The *Mir* space station, in orbit from 1986 to 2001, has provided numerous additional opportunities to conduct materials-science and biotechnology experiments in the microgravity environment. Single crystals of semiconductor materials with improved characteristics have been grown. A tungsten-aluminum alloy composite material that cannot be produced on Earth was obtained in space; the material is lightweight and possesses good strength and wear resistance. Other experiments have studied the convergence of cells in microgravity for the purpose of obtaining very pure biologically active substances to serve as a basis for producing unique vaccines. Biotechnical experiments, such as the purification of proteins using electrophoresis, are performed in installations that are reported to be much more productive than on Earth. A technological module called *Kristall* was launched to *Mir* in 1990, providing the space station with a miniature plant for space production of ultrapure crystals and medical preparations. Analysis of these experiments is ongoing.

These efforts will be expanded using facilities on the International Space Station. A number of the international partners have developed laboratories to take advantage of, and investigate the influence of, weightlessness. These facilities are directed at the growth of purified crystals to better define the functional components of their structures, processing of materials to obtain a uniformity not achieved on Earth, and expanding basic knowledge of the effects of gravity on structures and physical reactions. An example of these efforts is the Granada Crystallization Box experiment. The experiment container, launched in August 2001, allows observations of the effect of variable levels of convection on the growth of protein crystals. This information will be used to compare different methods of growing crystals and model crystal growth in the absence of gravity. *See* SPACE FLIGHT; SPACE PROCESSING; SPACE STATION.

Thora Waters Halstead; Charles Edwin Wade

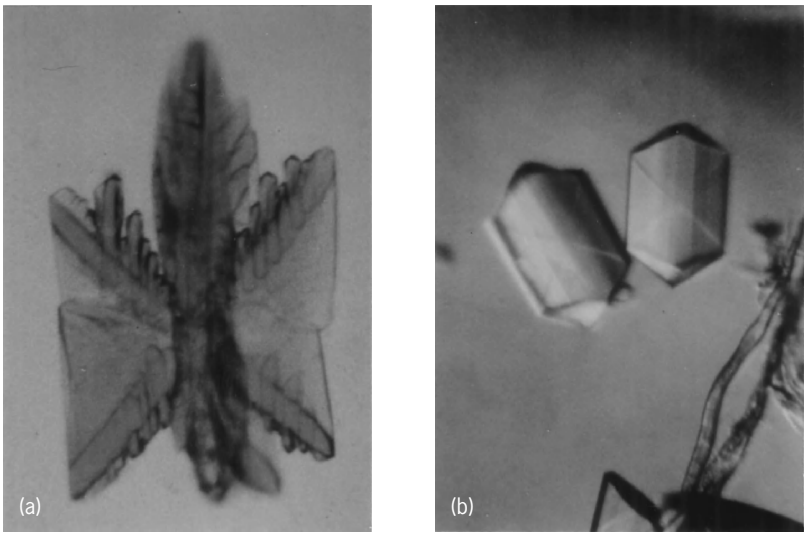


Fig. 4. Comparison of isocitrate lyase crystals (*a*) grown on Earth with typical dendritic morphology of Earth-grown crystals (*from P. C. Weber, Du Pont de Nemours & Co.*) and (*b*) grown on the space shuttle in September 1988 (*from L. J. DeLucas et al., Protein crystal growth in microgravity, Science 246:651–654, 1989*).

Bibliography. R. Berisio et al., Effects of microgravity on the crystal quality of a collagen-like polypeptide, *Acta Crystallog. D Biol. Crystallog.*, 56:55–61, 2000; T. G. Hammond et al., Gene expression in space, *Nat. Med.*, 5:359, 1999; Y. Kamotani, S. Ostrach, and A. Pline, Analysis of velocity data take in Surface Tension Driven Convection Experiment in microgravity, *Phys. Fluids*, 6:3601–3609, 1994; H. S. Lee, H. Merte, and F. P. Chiaramonte, The pool boiling curve in microgravity, *J. Thermophys. Heat Transfer*, 11:216–222, 1997; D. C. Walther, A. C. Fernandez-Pello, and D. L. Urban, Space shuttle based microgravity smoldering combustion experiments, *Combustion and Flame*, 1162:398–414, 1999.

Welded joint

The joining of two or more metallic components by introducing fused metal (welding rod) into a fillet between the components or by raising the temperature

of their surfaces or edges to the fusion temperature and applying pressure (flash welding).

Types. In a lap weld, the edges of a plate are lapped one over the other and the edge of one is welded to the surface of the other (**Fig. 1**).

In a butt weld, the edge of one plate is brought in line with the edge of a second plate and the joint is filled with welding metal or the two edges are resistance-heated and pressed together to fuse.

For a fillet weld, the edge of one plate is brought against the surface of another not in the same plane and welding metal is fused in the corner between the two plated, thus forming a fillet. The joint can be welded on one or both sides.

A weldment is a cast steel, forged steel, or machined steel component that is assembled to plates or structural steel shapes or other weldments by welding to form a machine part (**Fig. 2**).

A spot weld is an electrical-resistance lap weld wherein lapped surfaces are brought to high temperature by the resistance to a low-voltage, high-amperage current between two water-cooled electrodes (**Fig. 3**). This weld is used primarily for thin sheet stock.

A seam weld is similar in production to a spot weld except that the electrodes are rollers or wheels rotating at constant speed; they produce a long narrow weld. The butted edges of pipe or tubing can be welded by a modified seam weld.

Strength. Because welded joints are usually exposed to a complex stress pattern as a result of the high temperature gradients present when the weld is made, it is customary to design joints by use of arbitrary and simplified equations and generous safety factors.

The force F of direct loading, and consequently the stress S, is applied directly along or across a weld. The stress-force equation is then simply $F = SA$, in which A is the area of the plane of failure (**Fig. 4**).

For eccentric loading, the force F causes longitudinal and transverse forces of varying magnitudes along

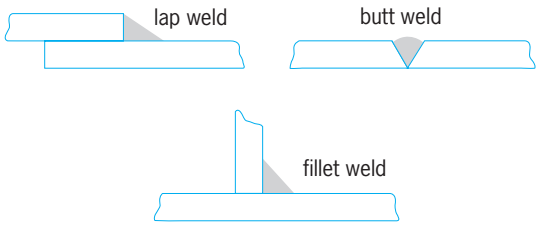


Fig. 1. Three types of welded joints.

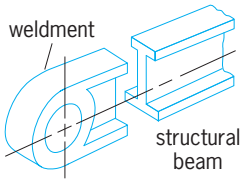


Fig. 2. Weldment is a preformed component designed to be welded to other structural or machine components.

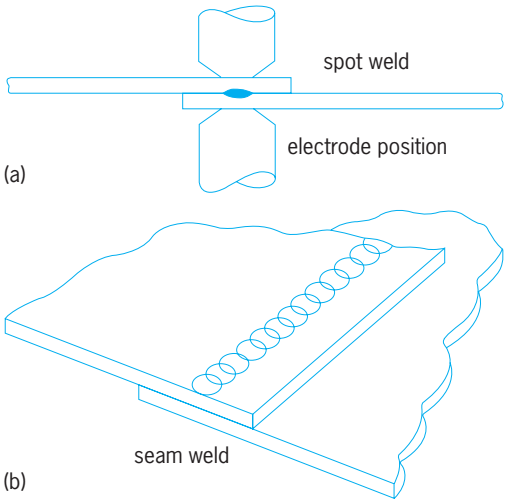


Fig. 3. Common electrical-resistance welds. (*a*) Spot weld, used mainly for thin sheet stock. (*b*) Seam weld, used to weld butted edges of pipe or tubing.

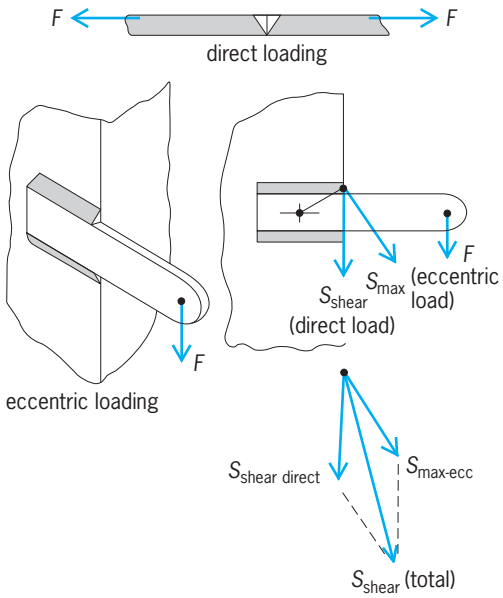


Fig. 4. Loading forces on a welded joint.

the weld. The stress is found by assuming that rotation occurs around the centroid of the welded area and that the shear stress due to the torque about the centroid is vectorially added to the direct shear caused by the force. Several points in the welded area must be checked to establish the maximum vectorial summation of the direct and eccentric stress. Maximum shear stress S_{max} is Fdr_{max}/J_{total}, where J_{total} is total polar moment of inertia of weld failure sections about the centroid, d is the distance from force F to the centroid, and r_{max} is the radius from the centroid to the farthest part of the weld. *See* STRUCTURAL CONNECTIONS; WELDING AND CUTTING OF MATERIALS.

L. Sigfred Linderoth, Jr.

Bibliography. A. D. Althouse et al., *Modern Welding*, 9th ed., 2000; T. B. Jefferson and D. T. Jefferson (eds.), *The Welding Encyclopedia*, 18th ed., 1988.

Welding and cutting of materials

Processes based on heat to join and sever metals. Welding and cutting are grouped together because, in many manufacturing operations, severing precedes welding and involves the same production personnel. Welding is one of the joining processes, others being riveting, bolting, gluing, and adhesive bonding. *See* BOLTED JOINT; WELDED JOINT.

The American Welding Society's definition of welding is "a metal-joining process wherein coalescence is produced by heating to suitable temperatures with or without the application of pressure, and with or without the use of filler metal." Brazing is defined as "a group of welding processes wherein coalescence is produced by heating to suitable temperature and by using a filler metal, having a liquidus above 800°F and below the solidus of the base metals. The filler metal is distributed between the closely fitted surfaces of the joint by capillary attraction." Soldering is similar in principle, except that the melting point of solder is below 800°F. The adhesion of solder depends not so much on alloying as on its keying into small irregularities in the surfaces to be joined. For comparison of metal joints. *See* BRAZING; JOINT (STRUCTURES); SOLDERING.

Cutting is one of the severing and material-shaping processes, some others being sawing, drilling, and planning. Thermal cutting is defined as a group of cutting processes wherein the severing or removing of metals is effected by melting or by the chemical reaction of oxygen with the metal at elevated temperatures. Welding and cutting are widely used in building ships, machinery, boilers, spacevehicles, structures, atomic reactors, aircraft, railroad cars, missiles, automobiles, buses and trailers, and pressure vessels, as well as in constructing piping and storage tanks of steel, stainless steel, aluminum, nickel, copper, lead, titanium, tantalum, and their alloys. For many products, welding is the only joining process that achieves the desired economy and properties, particularly leak-tightness. *See* TORCH.

Nearly all industrial welding involves fusion. The edges or surfaces to be welded are brought to the molten state. The liquid metal bridges the gap between the parts. After the source of welding heat has been removed, the liquid solidifies, thus joining or welding the parts together. The principal sources of heat for fusion welding are electric arc, electric resistance, flame, laser, and electron-beam. The flame in gas welding is provided by the combustion of a fuel gas.

Arc welding. The greatest volume of welding is done with arc welding processes.

Shielded metal-arc welding. This arc welding process is by far the most widely used of the various electric-arc welding processes. Like the other electric-arc welding processes, it employs the heat of the electric arc to bring the work to be welded and a consumable electrode to a molten state. The work is made part of an electric circuit known as the welding circuit. Arc welding with consumable electrodes is more widely practiced than welding with nonconsumable electrodes. A consumable electrode is melted continuously by the arc, one pole of which is the electrode, the other pole being the metal to be welded (**Fig. 1**). Generally, the consumable electrode is a wire of the same chemical composition as the work. The arc melts the electrode and some of the base metal to form a pool of weld metal, which after freezing becomes the weld. A nonconsumable electrode is tungsten or carbon, both of which have high melting points and are not consumed, except slowly by vaporization.

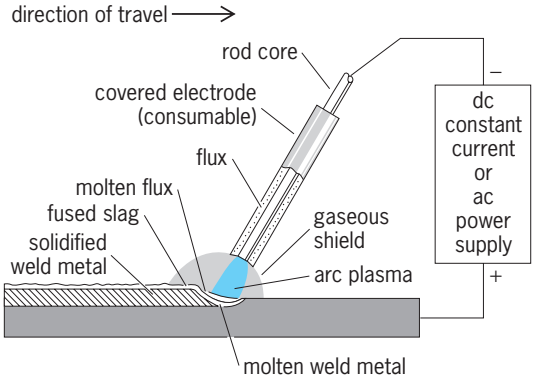


Fig. 1. Shielded metal-arc welding.

A covered electrode consists of a solid metal core; in some instances there are tubular-type cores which are covered with material serving electrical and metallurgical purposes. Electrically, the covering insulates the metal core from accidental contact with adjacent material during welding and provides an arc free from interruptions. Metallurgically, the covering may provide gas- and slag-forming ingredients to protect the weld from the air, and it may supply deoxidizers or alloying elements to produce sound welds having specified chemical composition. In addition, the covered (or coated) electrode provides scavengers to refine the grain structure of the weld metal. As the arc consumes the electrode, the operator manually feeds it into the weld.

About 6.35×10^8 lb (2.88×10^8 kg) of covered electrodes are manufactured every year with diameters ranging from $^1/_{16}$ to $^1/_4$ in. (1.6 to 6.4 mm) in stainless steel, $^5/_{64}$ to $^5/_{16}$ in. (2.0 to 7.9 mm) in mild steel, $^3/_{32}$ to $^5/_8$ in. (2.4 to 15.9 mm) in tool steel, and comparable diameters for the nonferrous alloys. The weight of the covering may be 10–50% of the weight of the covered electrode. For welding mild steel, covered electrodes are made in a number of types, which have been classified by the American Welding Society. Class E-6010 has a covering consisting principally of cellulose. The electrode can be used readily to weld flat, vertical, and overhead surfaces. The prefix E designates an electrode. In the specification for mild-steel-covered arc welding electrodes, the first two digits stand for a minimum tensile strength of the deposited metal in the as-welded condition in 1000 lb/in.^{2} (6.895 megapascals). The third digit stands for the welding position or positions in which the electrode will make a satisfactory deposit, and the last digit classes the performance

characteristics of the electrode and identifies the types of power sources on which it can be used. Class E-6012 electrodes have a thin covering high in rutile (titanium dioxide) and are adapted to welding across gaps up to $^1/_4$ in. (6.4 mm). Class E-6016 electrodes are the low-hydrogen class. The calcium carbonate in the covering evolves carbon dioxide during welding. The covering is nearly free from hydrogen, and the weld, therefore, is free from defects due to hydrogen. Class E-6024 electrodes have about 40% iron powder in the covering, which makes possible high rates of depositing weld metal. The rate of depositing steel weld metal from steel electrodes in any arc welding process ranges from 0.03 to 0.06 lb/(min)(100 A) [14 to 27 g/(min)(100 A)] of welding current.

Cored electrodes consist of a tube formed from strip and filled with slag-forming, arc-stabilizing, and alloying materials. For welding mild steel, the cored electrode is continuous. In automatic welding the arc is guided mechanically along the joint, whereas in semiautomatic welding the operator guides the arc manually by means of a torch. In some processes the arc is in air. In others a gas such as carbon dioxide protects the arc from the air.

Submerged-arc welding. This process is performed with a continuous electrode and a granular flux composed of silicates with or without deoxidizers and alloying elements. The flux is piled to a depth of $^1/_2$ to 2 in. (12.7 to 50.8 mm) along the joint. The arc melts some of the flux and is submerged in the liquid slag so produced. Although submerged-arc welding often uses current of about 40 nanometers, it can use as much as 400 nm, far above that usable in any other arc welding process. High currents enable the weld to penetrate more deeply below the surface of the work than in other arc welding processes. Welds of exceptionally good quality are produced in a wide range of metals greater than $^1/_{16}$ in. (1.6 mm) thick. Carbon, alloy, or stainless steels up to $^1/_2$ in. (12.7 mm) thick are safely welded in one pass, while thicker materials are multipass-welded. Weld-metal deposition rates (pounds of metal deposited in 1 h of operation), arc travel speeds, and weld completion rates are superior to any other process available today (**Fig. 2**).

Gas metal-arc welding. Inert-gas-shielded metal-arc welding with consumable, continuous electrodes (called GMAW for gas metal-arc welding) requires no flux and produces welds without a slag cover. The arc is in an atmosphere of argon or helium supplied at 5–100 ft^3/h (0.14–2.8 m^3/h) from the gas cup of the torch. Inert-gas shielding is particularly advantageous in welding reactive metals, such as titanium, which are susceptible to atmospheric contamination, and in welding metals, such as aluminum and stainless steel, which are susceptible to porosity (**Fig. 3**).

Carbon dioxide welding. Carbon dioxide welding with continuous electrodes is similar to gas metal-arc welding except that carbon dioxide (CO_2) is used as a low-cost shielding gas. The process is restricted to carbon and some low-alloy steels. To prevent porosity resulting from the formation of carbon monoxide by the reaction of the CO_2 with the carbon in the steel, electrodes for CO_2 welding contain deoxidizers such as manganese, silicon, aluminum, titanium, and chromium.

Nonconsumable arc welding electrodes are not deposited as part of the weld metal. The electrode is one pole of the arc, usually the electron-emitting (and therefore cooler) negative pole. The other pole is the work, which is melted. If additional metal is required to fill the joint, a filler wire is fed into the arc. Only tungsten and carbon have sufficiently high melting points to provide requisite electron emission at the high currents used in arc welding. Carbon electrodes seldom are used for welding because they vaporize more rapidly than tungsten.

Gas tungsten-arc welding. Inert-gas shielding is essential with tungsten electrodes; consequently the term "gas tungsten-arc welding" (GTAW) is used. A common electrode diameter is $^1/_8$ in. (3.2 mm) for 150 A. The process is adapted to welding thin material, $^1/_8$ in. (3.2 mm) thick and less, and to the root, or first pass, because the operator can control penetration more readily than with most other arc welding processes. The process is especially adapted for welding light-gage work requiring the utmost in quality or finish, because of the precise heat control possible and the ability of weld with or without filler metal (**Fig. 4**). Direct current (electrode negative) is used

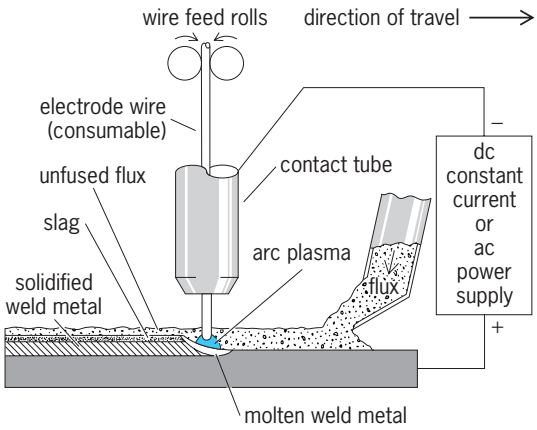


Fig. 2. Submerged-arc welding.

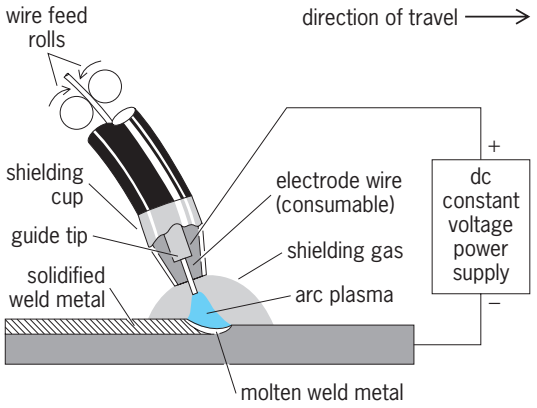


Fig. 3. Gas metal-arc welding.

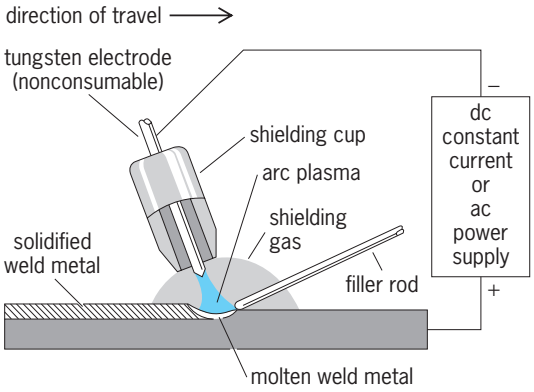


Fig. 4. Gas tungsten-arc welding.

in welding ferrous materials, copper, and nickel alloys. For aluminum and magnesium, alternating current is required because oxide film on the work is removed by reverse polarity (electrode positive). To prevent extinction of the arc as the electric potential passes through zero, a small high-frequency current (100,000 Hz) must be superimposed on the 60-Hz welding current.

Plasma-arc welding. This is a process which utilizes a plasma produced by the heat of a constricted electric-arc–gas mixture. Shielding gas may be an inert gas or a mixture of gases. Two arc modes are used to generate a plasma: transferred arc and nontransferred arc.

Plasma-arc welding resembles gas tungsten-arc welding in its use of an inert gas, but differs from it in the use of a constricting orifice. The transferred-arc mode is usually preferred. The advantages of plasma-arc welding have been observed primarily in material thicknesses greater than $^{3}/_{32}$ in. (2.4 mm). In such thicknesses, a significant difference is observed in the weld puddle; it is known as the keyhole effect. Surface tension causes the molten metal to flow around the keyhole to form the weld. This keyhole can be observed during the welding operation and is an indication of complete and uniform weld penetration. Another process, plasma-MIG, combines the features of plasma-arc and inert-gas metal-arc processes. The narrow arc is used for deep penetration welding of thick material or high-speed welding of thin plate. *See* ARGON.

Flux-cored arc welding. Flux-cored arc welding is somewhat like submerged arc and shielded metal-arc welding except that the flux is encased in a metal sheath instead of being laid over the wire. Although wire is automatically fed from a coil, the equipment is portable and more versatile than submerged-arc. The weld metal is shielded by the melted flux and by a gaseous medium, either externally supplied or evolved from the flux. The two principal variations in the process employ a gas-shielded electrode that requires additional shielding in the form of CO_2 around the arc and weld puddle, or a self-shielded electrode that generates its own shielding (no external shielding gas is supplied; **Fig. 5**).

The classification system for flux-cored electrodes follows the general pattern used in other filler metal classifications. In a typical designation, E70T-1, the prefix E indicates an electrode as in other electrode classification systems; the number 70 refers to the minimum as-welded tensile strength in 1000-lb/in.^{2} (6,895-MPa) units, the letter T indicates that the electrode is of tubular construction (is a flux-cored electrode), the suffix 1 places the electrode in a particular grouping based on the chemical composition of deposited weld metal, method of shielding, and adaptability of the electrode for single- or multiple-pass use.

Electroslag welding. This welding process is initiated much like the conventional submerged-arc process by starting an electric arc beneath a layer of granular welding flux. As soon as a sufficiently thick layer of hot molten slag is formed, all arc action stops and current passes from the electrode to the workpiece through the conductive slag. At this point, the process is truly electroslag welding.

Heat generated by the resistance to the current through the molten slag is sufficient to fuse the edges of the workpiece and the welding electrode. Since no arc exists, the welding action is quiet and spatter-free. The interior temperature of the bath is in the vicinity of 3500°F (1925°C). The liquid metal coming from the filler metal and the fused base metal collects in a pool beneath the slag bath and slowly solidifies to form the weld.

The wire electroslag process has been used for welding plates ranging in thickness from approximately $^{1}/_{2}$ to 20 in. (1.25 to 50 cm). The welding of greater thicknesses is technically feasible, but requires custom installations. With each electrode, the process will deposit from 25 to 45 lb (11 to 20 kg) of filler metal per hour. The diameter of the electrode wire generally used is $^{1}/_{8}$ in. (3.2 mm). Materials that have been joined by the electroslag welding process include mild and low-alloy carbon steels, stainless and high-alloy steels, and titanium.

Gas tungsten-arc spot welding. A combination of gas tungsten-arc and resistance spot welding has also been developed. Gas tungsten-arc spot welding is quite versatile and can be used to produce excellent-quality spot welds. Argon is used to shield

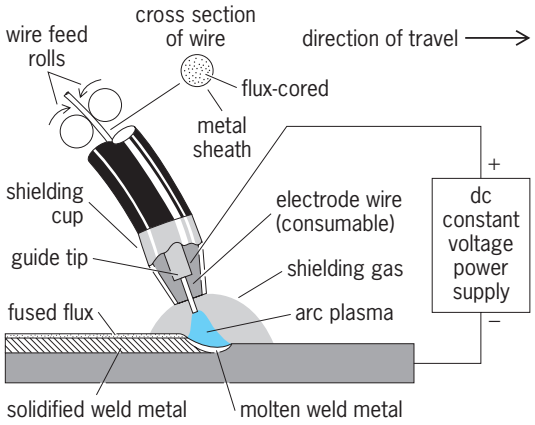


Fig. 5. Flux-cored wire welding.

the tungsten electrode and the face of the spot weld. On materials that require protection from air, the opposite side of the spot weld is shielded with either argon or helium. *See* ARC WELDING.

Resistance welding. Resistance welding processes are widely used in the manufacture of sheet metal assemblies, such as automobile bodies, aircraft, missiles, railroad cars, buses, and trailers. By definition, the required heat at the joints to be welded is generated by the resistance offered through the work parts to the relatively short-time flow of low-voltage, high-density electric current. Force is always applied before, during, and after applying current to assure a continuous electrical circuit and to forge the heated parts together.

In the most widely used processes—spot, seam, flash, and projection welding—the heat melts the surfaces to be welded. Every resistance weld involves a sequence of electrical energy and mechanical pressure. The sequence is provided by a control, which governs timing of both. For example, a control meters the 10 cycles of 60-Hz power at 10,000 A and also the electrode force at 500 lb (227 kg) required for spot welding two sheets of 0.04-in. (1-mm) steel. The welding operator merely presses the button that sets the control in operation (**Fig. 6**).

Spot welding. In spot welding, the sheets to be welded are held between electrodes of hard, high-conductivity copper alloy. They conduct heat away from the electrode-to-sheet contacts. The resistance heat at the sheet-to-sheet contact fuses the sheets together. Electrode pressure provides uniform sheet-to-sheet resistance and prevents expulsion of metal between the sheets. *See* SPOT WELDING.

Seam welding. Seam welding is a process for making overlapping spot welds by means of rotating electrode wheels. The joints can be made gas-tight and liquid-tight, as in tubing and stainless steel electric refrigerators.

Flash welding. In flash welding, the surfaces to be welded are held lightly in contact while a high current flows through the few contact points. Melting occurs instantaneously at these "bridges." Violent vaporization ensues, with expulsion of small particles of hot metal as visible flash. Flashing is continued until the surfaces are coated with a layer of molten metal. They are then squeezed together to form the weld. Flash welding is used to a considerable extent in the manufacture of automotive and aircraft products. The process is also used widely in the manufacture of household appliances, refrigerators, and farm implements. *See* FLASH WELDING.

Three other welding processes similar to flash welding should be mentioned: upset, upset butt and tube, and percussion welding. The greatest use of upset butt welding is in wire mills and in the welding of products made from wire. Equipment using the percussion welding process is found in several industries, mainly those in the electrical contact or component field.

Projection welding. In projection welding, a projection may be embossed on a sheet of metal. It may be produced on a solid piece of metal by machining, or it may be produced on an edge in a punch press in several ways. The height may be anywhere from a few thousandths of an inch (2.5 cm) to $^1/_8$ in. (3.2 cm) or more, depending upon the job. The purpose of projections is to localize the current and pressure at predetermined points. In this modification of the spot welding process, the concentration of the welding current is determined by the preparation of the workpieces rather than by the size and shape of the electrodes. The major portion of the heat tends to develop in the part bearing the projections during the welding operation. Cross-wire welding, which is used extensively in making wire mesh and other forms of wire products, is a form of projection welding. *See* RESISTANCE WELDING.

Gas welding. A gas flame is a less concentrated and lower-temperature source of heat than an arc. For this reason, a welding torch is often used for brazing, for welding thin material, and for applications requiring low gradients of temperature to avoid cracking, as in welding cast iron. The oxyacetylene flame, using a mixture of equal volumes of oxygen and acetylene (C_2H_2), has a higher temperature than any other commercially available fuel gas combination. High flame temperature is essential for the rapid localized heating required in welding. Several sizes and types of torch tips are available to secure flames of different intensities and dispersions. The ratio of oxygen to acetylene governs the chemical action of the flame. If the ratio is higher than unity, for example 1.1:1, the flame is oxidizing and is useful for preventing hydrogen porosity in brass. With a ratio below unity, 0.95:1, an acetylene feather appears in the flame, carbon is picked up by molten steel, and mill scale on the surface of steel is reduced. A neutral flame with a ratio of 1:1 is used for steel.

Flux is not used in welding steel but is necessary for gas-welding nonferrous metals. The use of other fuel gases, such as hydrogen, is used in the welding of metals of low melting point, such as lead, aluminum, and magnesium.

Electron beam welding. Electron beam welding is a process wherein coalescence is produced by heat

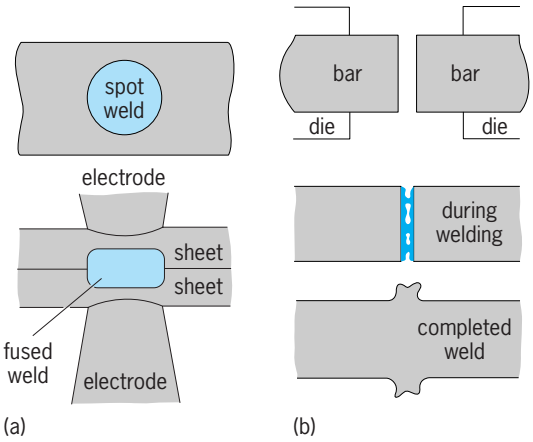


Fig. 6. Examples of resistance welding. (*a*) Spot welding (*b*) Flash welding.

obtained from a concentrated beam composed primarily of high-velocity electrons impinging upon the surfaces to be joined. Electron beam welding equipment, which utilizes a heat source of electrons accelerated by an electric field to extremely high speeds and focused to a sharp beam by electrostatic or electromagnetic fields, is used for welding a wide range of metals in thicknesses ranging from foils to extremely thick sections. The welding equipment is manufactured for welding in high vacuum of thicknesses of 0.1 to 0.01 micrometer, in medium vacuum of 50 to 300 μm, and out of vacuum in an air or inert-gas atmosphere. The basic welding equipment is designed to operate in one of two voltage ranges. So-called low-voltage welding units operate at 15 to 60 kV, while high-voltage welders operate at 100 to 200 kV. A weld of very high purity may be obtained by this technique. Of all welding techniques, electron beam welding seems most suited for use in outer space. *See* ELECTRON MOTION IN VACUUM; VACUUM METALLURGY.

Mel M. Schwartz

Laser welding. In the continuous laser beam welding process, the laser beam is focused on the workpiece surface, and at the point of focus the metal vaporizes under the beam and a vapor column extending deep into the base metal is created. The vapor column is surrounded by a liquid pool in equilibrium with the vapor (**Fig. 7**). By traversing the path to be joined with the focused beam, autogenous deep-penetration fusion welds are created (Fig. 7). Depth-to-width ratios for these welds generally exceed four to one, and are often as great as ten to one.

Low-power lasers such as pulsed ruby or pulsed yttrium-aluminum-garnet (YAG) are suitable for welding thin-gage materials such as electronic components. These lasers can produce millisecond pulses of up to 100 kW for spot welding.

Deep-penetration autogenous welds are formed primarily with continuous multikilowatt gas laser systems, the most highly developed and widely used of which is the carbon dioxide laser, operating in the far-infrared region at a wavelength of 10,600 nm. These lasers, when focused, are capable of producing power densities adequate to initiate deep penetration, a phenomenon which was previously possible only with electron beams.

The laser welding process is ideally suited to automation, since the beam constitutes a clean, remote heat source which can be readily shaped and directed by using reflective optics. A vacuum environment is not required, and a choice of shielding gasses and gas shield configurations may be adapted to specific welding situations. Unlike the somewhat similar electron beam welding process, laser welding does not generate x-rays.

The laser welding process has been shown to be suitable for a variety of metals and alloys, including ferrous materials, aluminum alloys, titanium alloys, lead, copper and copper-nickel, and most superalloys and refractory metals. Laser welds have been fabricated in these materials, which are generally ductile and porosity-free, and have average tensile strengths equivalent to or exceeding those of the base metal. High weld strengths are the result of high cooling rates, which result from the low energy input per unit length of weld which is characteristic of the laser. These high cooling rates often are advantageous in that they produce desirable structures in the welds.

Laser beam welding typically requires good joint fit-up, and is a high-speed process ideally suited to automation. As a result of the generally high cost of laser welding equipment in excess of 1-kW power level, the process is most successfully applied to high-volume applications or to critical applications requiring unique weld characteristics. *See* LASER WELDING.

Edward M. Breinan

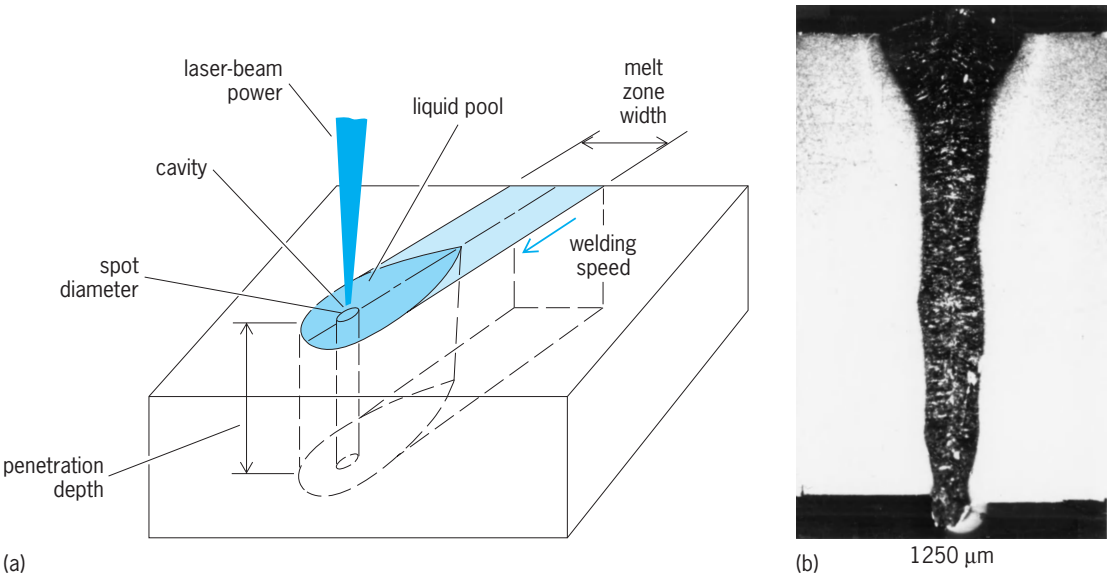


Fig. 7. Deep-penetration welding characteristics. (*a*) Schematic diagram of the deep-penetration laser welding process. The laser beam creates a vapor column (deep-penetration cavity) and deposits energy through the thickness of the workpiece. (*b*) Typical weld.

Inertia welding. Inertia welding is one type of friction welding process which utilizes the frictional heat generated at a pair of rubbing surfaces to raise the temperature to such a degree that the two parts can be forged together and form a solid weld. The process is called inertia welding because the energy required for making the weld comes from a rotating flywheel system built into the head stock of the machine. Somewhat similar to an engine lathe, the inertia welding machine also has a tail stock where one side of the workpiece is held and can be pushed forward and retracted by a hydraulic ram.

During inertia welding, one part of the workpiece held in the spindle chuck of the head stock is accelerated rapidly to a preset spindle speed. The drive power is then cut off while the nonrotating workpiece on the tail stock is pushed against the rotating part under a large thrust force. The friction at the interface brings the spindle to full stop in a matter of seconds. In the meantime, the kinetic energy stored in the spindle-flywheel system is converted into frictional heat which causes the temperature at the weld to rise quickly so that a forge weld can be formed. Because of the high pressure, the original materials at the rubbing surfaces are first softened and then squeezed out of the interface to form a flash; thus, the process brings the nascent sublayers of the material close together and results in a metallurgical bond. The weld is formed at solid state without melting; it has a finer grain structure and narrower heat-affected zone as compared with any fusion welding process, for example, electric arc welding. This often results in higher strength at the weld than at the base material.

Inertia welding is commonly used to join dissimilar materials which cannot be achieved easily or economically by any other means. Material pairs of large varieties can be welded by this process, for example, steels to aluminum alloys or copper, copper to titanium or zirconium alloys, and many others. In addition to these advantages, inertia welding machines can be readily automated for mass production. The automotive and oil drill industries are the major users of the process. Typical applications include engine exhaust valves, shock absorber parts, universal joints, axle housings, hydraulic cylinder components, and oil drill pipes. The process is also applied to weld stainless steel to plain carbon steel for outboard motor shafts, and gas turbine parts in the aircraft industry. *See* INERTIA WELDING. K. K. Wang

Other fusion welding processes. Among other fusion welding processes are high-frequency, thermit, stud, and plastic welding.

High-frequency welding may be divided into two categories: (1) high-frequency resistance welding, in which the high-frequency current is introduced into the work by direct electrical contact; and (2) high-frequency induction welding, in which the high-frequency current is induced in the work by an inductor coil, without electrical contact. The proximity effect concentrates the current in the edges to be joined, as in the manufacture of metal tubes.

In thermit welding, the joint is produced by heating with superheated liquid metal and slag resulting from a chemical reaction between a metal oxide and aluminum, with or without the application of pressure. Filler metal, when used, is obtained from the liquid metal.

There are two basic types of stud welding; arc stud welding and capacitor-discharge stud welding. Arc stud welding, still the more widely used of the two basic stud welding processes, is similar in many respects to manual shielded metal-arc welding. The heat necessary for end welding of studs is developed by passage of current through an arc from the stud (electrode) to the plate (work) to which the stud is to be welded. Capacitor-discharge stud welding derives its heat from an arc produced by a rapid discharge of stored electrical energy with pressure applied during or immediately following the electrical discharge.

There are several different ways of joining plastics to each other: hot-gas welding, friction welding, heated tool welding, and ultrasonic welding. Welding is a convenient method of joining most thermoplastic materials. The procedure is similar to that followed in metal welding when the oxyacetylene process is used. The strength of the weld, however, may vary according to the type of plastic material. The strongest joints are usually obtained from plastics with the highest degree of polymerization.

Diffusion bonding. Diffusion bonding (also known as diffusion welding) is a solid-state process in which joining is accomplished without a liquid interface (brazing) or the creation of a cast product via melting and resolidification (welding). Diffusion bonding produces solid-state coalescence between two materials under the following conditions: (1) joining occurs at a temperature below the melting point of the materials to be joined; (2) coalescence of contacting surfaces is produced with loads below those that would cause macroscopic deformation to the part; and (3) a bonding aid can be used, such as an interface foil or coating, to either facilitate bonding or prevent the creation of brittle phases between dissimilar materials.

Diffusion bonding facilitates the joining of materials to produce components with no abrupt discontinuity in the microstructure and with a minimum of deformation. The process is generally considered complete when cavities fully close at the closely fitting (faying) surfaces.

Formation of a diffusion weld occurs in three stages: initial contact of the interfaces; a time-dependent deformation of the same interfaces, further establishing intimate interfacial contact; and diffusion-controlled elimination of the original interface. The final step may occur as the result of grain growth across the interface, the solution or dispersion of interfacial contaminant, or a simple diffusion of atoms along or across the original interface (**Fig. 8**).

Although not a part of the mechanism of diffusion bonding, the use of intermediate materials is of considerable importance. Intermediate materials can be used to promote diffusion at much lower temperatures than for self welding, promote plastic flow and

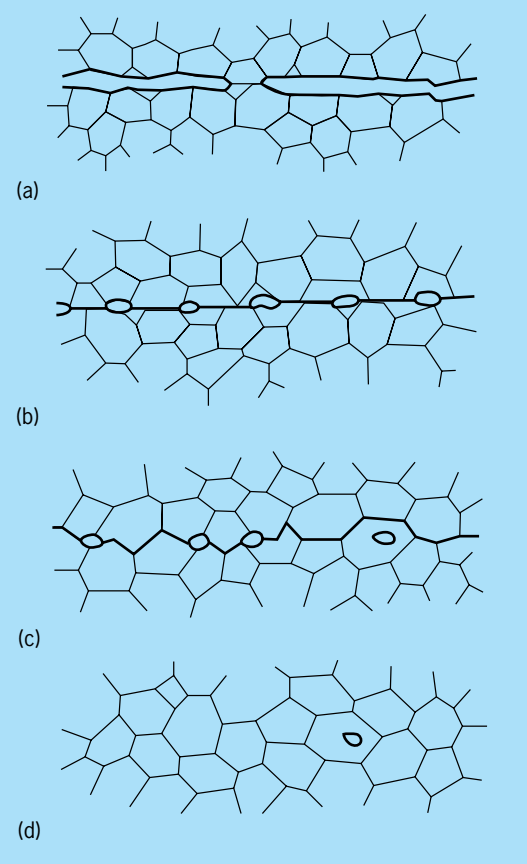


Fig. 8. Sequence of metallurgical stages in the diffusion bonding process. (*a*) Initial contact: limited to a few asperities (room temperature). (*b*) First stage: deformation of surface asperities by plastic flow and creep. (*c*) Second stage: grain boundary diffusion of atoms to the voids and grain boundary migration. (*d*) Third stage: volume diffusion of atoms to the voids.

surface conformance at lower pressures, prevent the formation of intermetallic compounds, and provide clean surfaces. The intermediate layer also permits welding in very short times. Intermediate or activating agent materials are commonly used in the form of foil, electroplates, or vapor deposits.

In practice, oxide-free conditions exist for only a limited number of materials. Accordingly, the properties of real surfaces limit and impede the extent of diffusion bonding. Titanium alloys as a group are the most notable exception. At temperatures greater than 1560°F (850°C), the titanium alloy can readily dissolve minor amounts of adsorbed gases and thin surface oxide films and diffuse them away from the bonding surfaces so that they will not impede the formation of the required metallic bonds across the bond interface.

Joint designs for producing parts that utilize intermediate diffusion aids to produce a diffusion weld are similar to those used for conventional brazing. The fundamental butt and lap joints with modifications such as scarf joints are used universally.

Materials. A number of materials can be joined by diffusion bonding; some are easier to diffusion-weld than others.

Aluminum and its alloys are difficult to join by diffusion bonding because of the tightly adhering oxide film. When the oxide is removed, the process can take place at temperatures of 849–1000°F (454–538°C) for up to 4 h. The time-temperature requirements can be considerably reduced by means of a diffusion aid such as silver or copper.

Other metals that can be joined by this process include beryllium and its alloys, copper and its alloys, heat-resistant cobalt or nickel alloys, various steels, columbium and its alloys, tantalum and its alloys, titanium and its alloys, and zirconium and its alloys.

Materials other than metals that can be joined by diffusion bonding include a number of composites, composite-metal combinations, and combinations of dissimilar metals and ceramics (such as ferrous and nonferrous materials).

Applications. Many industries have taken advantage of the benefits of the diffusion-bonding process, especially the aircraft industry, where the interest in titanium alloys is significant. Another application is in the refrigeration industry, where aluminum tubing can be substituted for copper in refrigerant line sets and the aluminum tubing subsequently joined to a brass end fitting. The metallurgical weld in this case is created without melting either the brass or the aluminum.

Mel M. Schwartz

Transient liquid-phase bonding. Compared to diffusion bonding, transient liquid-phase (TLP) bonding (also know as diffusion brazing) has the advantage of being much more tolerant of the presence of surface oxide layers on the substrates. It is also suitable for a wider range of geometries than diffusion bonding. However, unlike diffusion bonding, an interlayer is always required. As a result, the bond can never have exactly the same composition as the substrates. In some cases, the bonding times required to complete isothermal solidification and subsequent homogenization of TLP-bonded joints can be very long.

Transient liquid-phase bonding relies on the use of an interlayer, which is common with the brazing process. When heated to a suitable temperature, the interlayer forms a liquid between the materials to be joined (commonly known as the substrates or workpieces). In brazing, the composition of the interlayer is selected so that the liquid is formed at a temperature below the melting point of the substrates. This allows a wide variety of materials and geometries to be joined. Unfortunately, the use of a relatively low-melting-temperature interlayer makes conventional braze joints unsuitable for use in applications that involve exposure to high temperatures. In contrast, compositional changes that occur during transient liquid-phase bonding can result in a bond with a structure and properties very close to those of the substrates as in welding.

Process. The diffusion brazing process proceeds as shown in **Fig. 9** for a joint between two substrates consisting of some metal A. The substrates are joined using an interlayer consisting of metal A plus another metal B (Fig. 9*a*). Addition of B serves to reduce the melting point of the interlayer below that

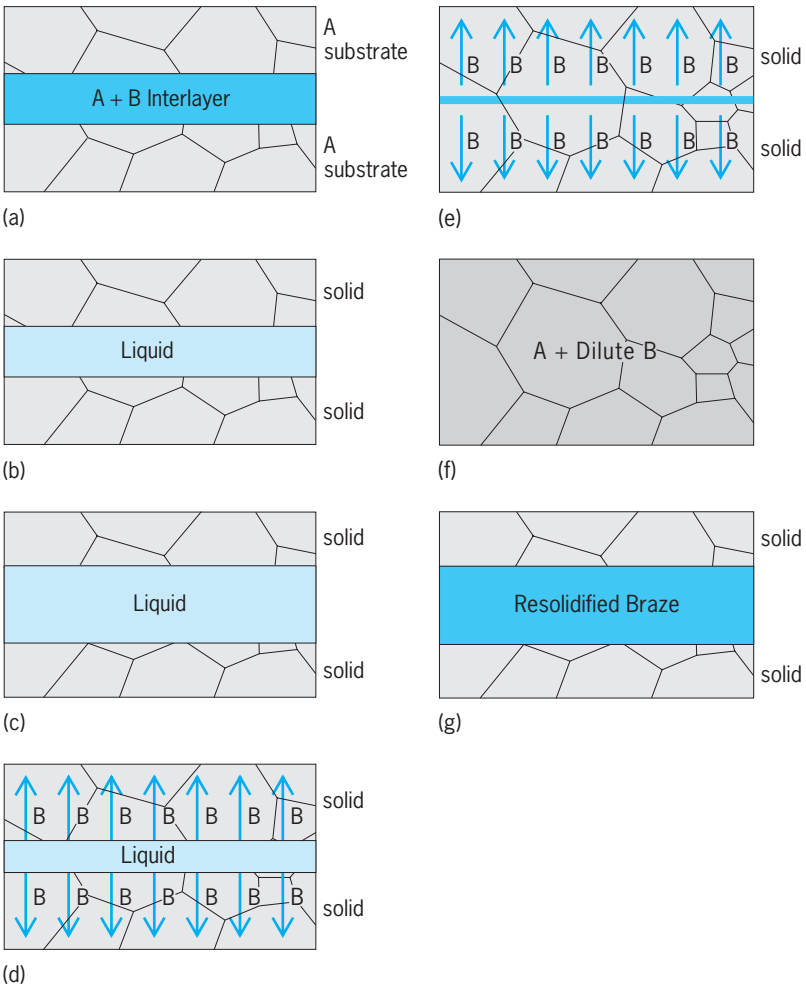


Fig. 9. Diffusing brazing process, showing the steps involved in the formation of a joint by transient liquid-phase bonding. (*a*) Initial condition. (*b*) Melting of interlayer. (*c*) Dissolution. (*d*) Isothermal Solidification. (*e*) Homogenization. (*f*) Final joint; diffusion brazing process, showing the steps involved in the formation of a joint by transient liquid phase bonding the crystals (grains) making up the substrate material extend across the bond. In comparison, (*g*) a conventional braze joint has gone through steps *a*, *b*, and *c*, and has then cooled.

of the substrates. The process temperature is chosen so that the interlayer melts but the substrates remain solid (Fig. 9*b*). After the interlayer melts, there is some tendency for the substrates to dissolve into the liquid and so the liquid initially becomes wider (Fig. 9*c*). Subsequently, atoms of metal B diffuse into the substrates (Fig. 9*d*). Because metal B reduces the melting point of the interlayer, the joint gradually resolidifies, as this metal diffuses away into the substrates (Fig. 9*d*). Normally, liquids resolidify only when cooled, but the compositional changes occurring during bonding allow resolidification of the joint while holding at a constant temperature. This is known as isothermal solidification. After isothermal solidification is completed, atoms of metal B continue to diffuse (Fig. 9*e*), and the result can be a homogenized joint that is almost identical to the substrates (Fig. 9*f*) and has very similar properties (unlike a conventional braze joint, as shown in{\uppercase{break}} Fig. 9*g*). Note that there is also a variant of the transient liquid-phase bonding process in which the interlayer starts as a pure metal and forms a liquid by reaction with the substrates.

Materials and applications. Transient liquid-phase bonding is used for both manufacturing and repairing components for high-temperature use, especially in gas turbines (for both aircraft engines and electric power generation). Mostly, the substrates involved are nickel-base alloys, and boron is often used to reduce the melting temperature of the interlayer.

William F. Gale

Explosive welding. Explosive welding is an important metal-joining technique because it can handle unique joining problems and large areas. Also known as explosive bonding, explosive welding is accomplished by a high-velocity oblique impact between two metals. The impact must have sufficient energy to cause the colliding metal surfaces to flow hydrodynamically when they intimately contact one another in order to promote solid-state bonding. Oblique impact is important because conservation of momentum allows for a reentrant jetting action that is due to hydrodynamic flow of the faying metal surfaces. The jet is ejected outward from the collision apex between the metals and produces a cleaning action by scarfing or effacing the metal surfaces. The resulting virgin metal surfaces are then compressed together under high pressure from the explosion, which promotes atomistic bonding.

The controlled application of the enormous power generated by detonating explosives permits cleaning of metal surfaces of contaminants such as films of oxide and nitride, oil, or adsorbed gases, and immediate pressing these thoroughly clean surfaces together at pressures of several million newtons per square meter. Further, the relatively small amount of energy involved with the process gives rise to minimum melt accumulation at the metal-metal interface. These are ideal conditions for effective welding.

Materials. Combinations of metals such as aluminum to steel and titanium to steel, which form brittle intermetallics when exposed to the elevated temperatures of conventional welding methods, are readily joined by explosive welding. The only metals that cannot be welded by the process are those too brittle to withstand the impact of the explosive impulse. In general, any metal or alloy possessing sufficient ductility and impact resistance such that it will not fracture as a result of the dynamic forces involved may be explosively welded. Cladding metals with elongation as low as 5% have been explosively welded. *See* CLADDING.

Process. To achieve a metallurgical bond between two similar or dissimilar pieces of metal, the atomic structure of the first metal must be brought sufficiently close to the atomic structure of the second metal that the cohesive atomic forces of each array of atoms can effectively act on the other. This normally requires an intimacy of contact on the order of an interatomic spacing. Because metal and metal-alloy surfaces in a normal atmospheric environment are always covered with films of oxides, nitrides, and adsorbed gases, the surfaces of the mating metal plates, even under very high direct pressure, will not bond metallurgically because sufficient intimacy of contact cannot be established.

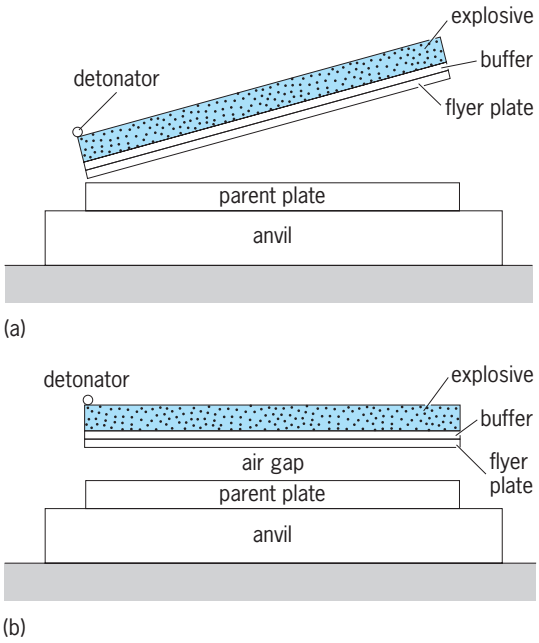


Fig. 10. Explosive welding arrangements. (*a*) Inclined. (*b*) Parallel.

The removal or effacement of surface films and the establishment of the necessary intimacy of contact must be achieved in any metallurgical bonding process. Fusion welding, brazing, and soldering are examples of joining processes that effectively disrupt surface films by dissolution or melting, and they establish the necessary intimacy of contact by wetting of the molten phase. However, processes such as friction welding and ultrasonic welding disrupt surface films by severe localized deformation of the mating surfaces while the surfaces are maintained in intimate contact by high pressure.

The process is basically simple. Typical arrangements are inclined or parallel (**Fig. 10**). The prime metal is oriented to the backer metal, with an air gap separating them. The explosive (in plastic, liquid, or granular form) is placed uniformly over the prime metal. The base of the backer rests on an anvil appropriate to the thickness of the backer and to the allowable distortion of the final product.

Applications. The widest application of explosive welding is in the area of cladding and welding of flat plate and sheet, which involves the simplest and most straightforward procedures. The major commercial application is the sheet and plate composite; the total thickness of sheet and plate that have been welded ranges from a few mils to over a foot, with cladding metal thickness ranging from 0.03 mm to more than 38.1 mm. The sizes of sheet and plate explosively welded range from small coupons to areas as large as 10 ft × 26 ft (3 m × 8 m).

Explosively welded products utilized commercially include clad tubular products with either internal or external clads. An example is the use of titanium–stainless steel transition joints in the Apollo lunar excursion module.

Welding metallurgy. The effects of welding on a product often govern its suitability for service. The weld metal, the zone close to the weld (called the heat-affected zone), and the thermal stresses and strains caused by welding may combine to yield a product different in strength, ductility, leak-tightness, and dimensional stability from that of the unwelded base metal. Weld metal frequently is compared with cast metal, with which it shares the characteristic of freezing in a columnar dendritic pattern in the direction opposite to the direction of withdrawal of heat. Unlike cast metal, weld metal usually has a free surface exposed to the air, and it is free from shrinkage cavities. Weld metal rapidly freezes compared with most castings and is comparatively fine-grained and free from chemical segregation. Porosity must be guarded against. It is a result of evolution of gas, such as nitrogen from steel, during freezing of the gas-saturated weld pool. Porosity also may result from a chemical reaction in the weld metal, such as the reaction between carbon and oxygen in liquid steel. Finally, porosity may occur as a result of inadequate inert-gas coverage during welding because of poor fixturing or poor welding conditions and procedures. Hot cracking of the weld metal during the last stages of solidification may be prevented either by fixturing to avoid strain during freezing or by adding suitable alloying elements to the weld, as in the addition of manganese to steel weld metal to combine with sulfur.

The heat-affected zone may be softer than the base metal, especially in alloys that have been cold-rolled or heat-treated for high strength. In some instances, for example the welding of annealed air-hardening steels, the heat-affected zone may be hard and brittle. Preheating and postwelding heat treatment may be required to avoid brittleness or restore strength. To achieve higher notch-impact value in steel weld metal than can be attained in a single pass, two or more passes must be deposited to take advantage of the grain-refining (hence toughening) effect of each pass on the heat-affected zone of the preceding pass. Welding heat causes distortion, which can be minimized by suitable choice of fixturing, joint design, and sequence of welding. The welded part also contains internal stresses, often up to yield strength locally in the vicinity of the weld. These may be reduced after welding by stress-relief heat treatment, which entails, in the case of steel, heating the welded structure to 1200°F (650°C) and cooling slowly. *See* HEAT TREATMENT (METALLURGY); METALLURGY.

Mel M. Schwartz

Microstructure modeling. The performance of a weld depends on its microstructure—that is, the spatial arrangement of its phases. The crystal structures or elemental compositions observed in the microstructure differ in the various regions of the welded structure, such as the base metal, the heat-affected zone (HAZ), and the fusion (melted metal) zone (FZ). During a typical welding process, microstructure evolution is affected largely by the rapid increase and decrease in temperature, which varies from region to region in the weld as shown in **Fig. 11**. These

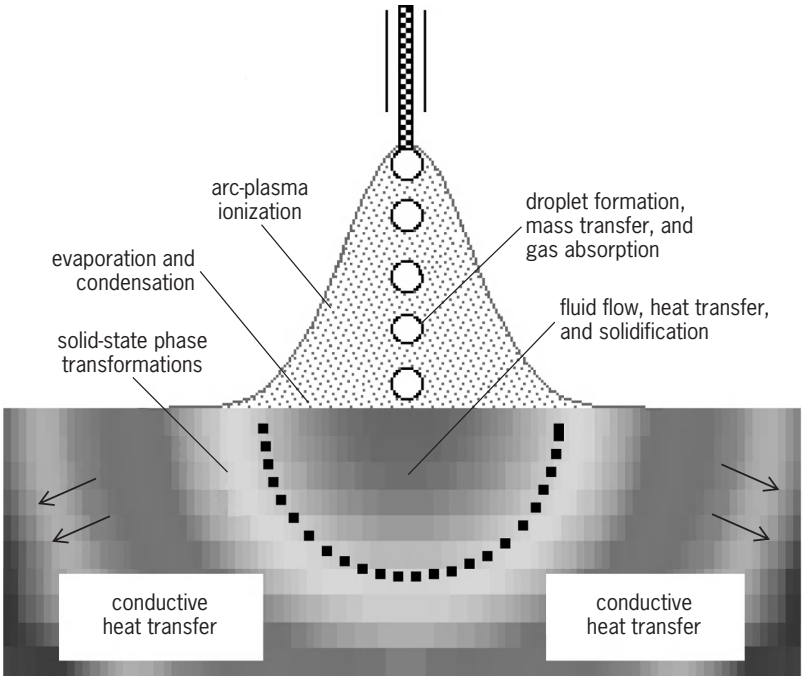


Fig. 11. Overview of physical processes in a typical arc welding process. The gray tints indicate temperature gradations.

temperature variations lead to severe microstructural gradients over a distance on the order of a few millimeters or less. *See* CRYSTAL STRUCTURE; METALLOGRAPHY.

Research is under way to describe these effects using computational models and experimental techniques to determine the best welding processes and process parameters for a given material. This is not a trivial task, considering the various physical processes that occur during welding (Fig. 11). The models are used to describe the microstructure evolution in the heat-affected zone and fusion (melted metal) zone for a given weld cooling rate and composition. Weld modeling also relies on thermal modeling efforts to relate the microstructure evolution to welding process effects. *See* MODEL THEORY.

Heat-affected zone modeling. Changes in the heat-affected zone, which reaches temperatures just below the melting point of the material, are mainly due to the response of the material to the thermal effects of solid-state phase transformations. If the material is a single-phase material, the grains in the heat-affected zone will grow as shown in **Fig. 12***a*. This may lead to a reduction in strength of the heat-affected zone. Grain growth kinetics as a function of weld thermal cycles can be described using analytical models or Monte Carlo simulations. *See* MONTE CARLO METHOD.

If the microstructure contains many phases, the changes may involve transformation from one phase to another based on composition and temperature gradients. These changes are closely linked to the thermodynamic stability of the phases and the diffusivity of the various elements in the phases.

Stainless steel that has both ferrite (body-centered cubic crystal structure) and austenite (face-centered cubic structure) phases undergoes changes as shown

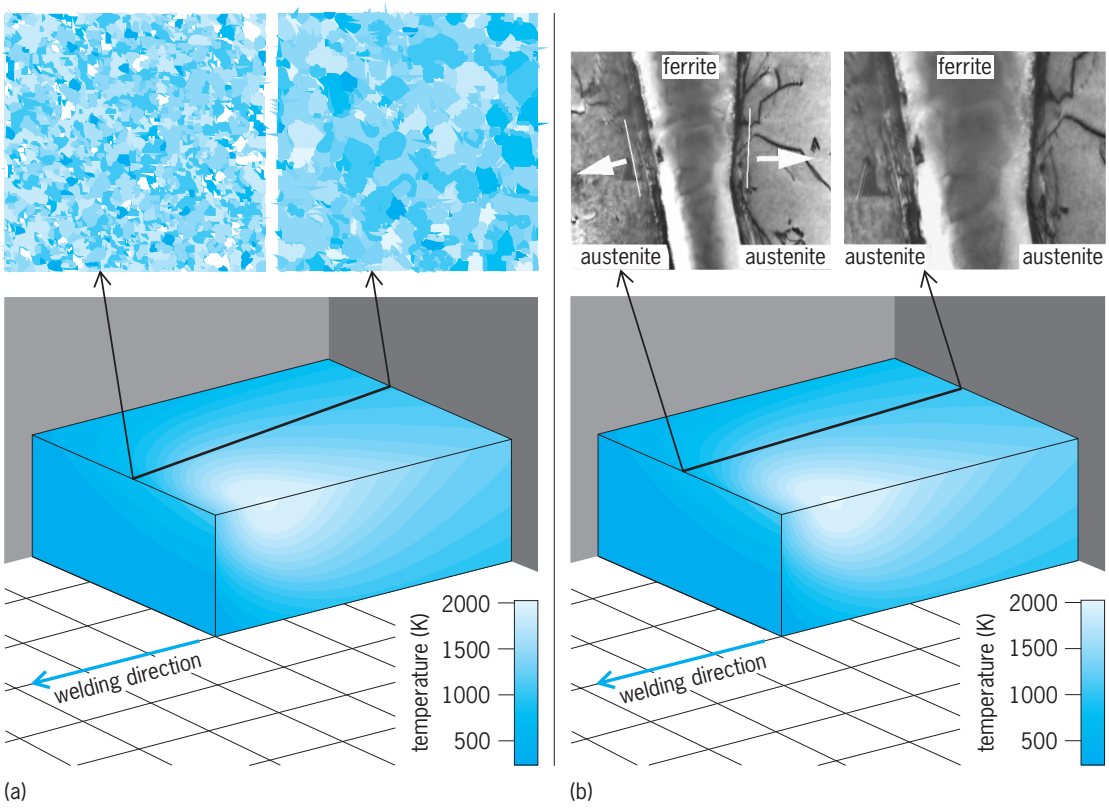


Fig. 12. Schematic representation of (*a*) grain growth in a single-phase material (colors indicate grains with different orientations) and (*b*) phase growth in a two-phase material (arrows in the multiphase materials show the direction of the boundary movement).

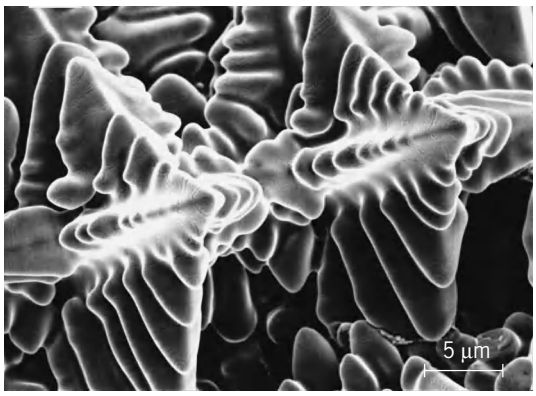


Fig. 13. Scanning electron micrograph showing a dendritic microstructure in a nickel-base superalloy weld.

in Fig. 12*b*. As the temperature increases, the stability of ferrite in the weld increases, and the ferrite grows into the austenite, leading to a larger fraction of ferrite. This growth involves diffusion of chromium, nickel, and carbon in both the phases. For the other compositions or thermal cycles of stainless steel, the ferrite may dissolve and the austenite may grow during welding. These complex changes may occur during weld cooling. The effects during heating and cooling do not cancel each other. The above changes may be beneficial or deleterious depending upon the service requirements of the welds. *See* STAINLESS STEEL.

Some of the solid-solid transformations in the heat-affected zone that occur in low-alloy (carbon) steel without diffusion (displacive transformations) can be modeled with thermodynamics and analytical expressions for the rate of the interface motion.

Fusion-zone modeling. Modeling the fusion zone is complicated by the need for a description of liquid to solid transformation (solidification), as well as the subsequent solid-state transformation. Some of the microstructural details in the fusion zone that are important are grain structure, grain morphology, partitioning of elements between liquid and solid, selection of phases in multiphase alloys, and physical defects such as cracks. During weld solidification, the solid phase grows epitaxially from the grains at the fusion line. The solidification substructure may be in the form of cells or dendrites, depending upon the thermal and compositional conditions near the solid-liquid interface (**Fig. 13**). With the growth of these dendrites, the alloying elements partition to the liquid, according to their thermodynamic properties. As the solidification continues, the partitioning of elements becomes more severe, leading possibly to nucleation of new grains in the liquid metal. At the end of solidification, cooling to room temperature results in other solid-state transformations. Modeling the fusion zone is complex, because in addition to solidification and solid-state transformation, the influence that solidification has on subsequent solid-state transformations must be considered. *See* CRACKING; EPITAXIAL STRUCTURES.

Grain structure evolution in single-phase materials has been modeled using Monte Carlo techniques, as well as some of the new modeling methods that use thermodynamic calculations to describe the solidification characteristics of multicomponent and multiphase alloys. For example, the fusion zone of stainless steel may contain either the austenite phase or the ferrite phase or a mixture of both, depending upon the composition and the cooling rate. The presence of ferrite in welds can be beneficial to avoid cracking but may be deleterious for corrosion resistance. Therefore, there is a need to develop models to describe the competition between these phases during welding. The relative thermodynamic stability between the ferrite and austenite phase can be correlated to the amount of ferrite that remains at room temperature. In terms of composition, adding chromium to stainless steel results in more ferrite formation, and adding nickel to stainless steel results in more austenite formation in welds. *See* CORROSION.

Thermodynamic and kinetic models are increasingly used to describe microstructure evolution in welds and in integrated models that describe the overall performance of the welds.

S. S. Babu; J. M. Vitek; S. A. David

Thermal cutting. Thermal cutting is used for a wide variety of work of the heavier type, such as structural work and shipbuilding, where the highest degree of dimensional accuracy is unimportant. Thermal cutting can be done manually or automatically, and in either case the torch can be brought to the work. The cut can be straight-line, curved, or beveled. The principal cutting process is oxygen cutting, which can be used for most mild steels. The cutting torch tip has a central orifice for oxygen and peripheral orifices for the preheat flame (**Fig. 14**). The flame preheats the steel to the ignition temperature, whereupon the oxygen converts the steel to iron oxide, which issues from the bottom of the cut as a molten spray. An oxygen cutting machine can cut a 1-in. (25.4-mm) steel plate at a speed of 18 in./min (46 cm/min), using an oxygen orifice of 0.0595-in. (1.51-mm) diameter and an oxygen pressure of 35 lb/in.^{2} (240 kilopascals) and consuming oxygen at the rate of 150 ft^3/h (4.25 m^3/h). Oxygen lancing is similar to cutting

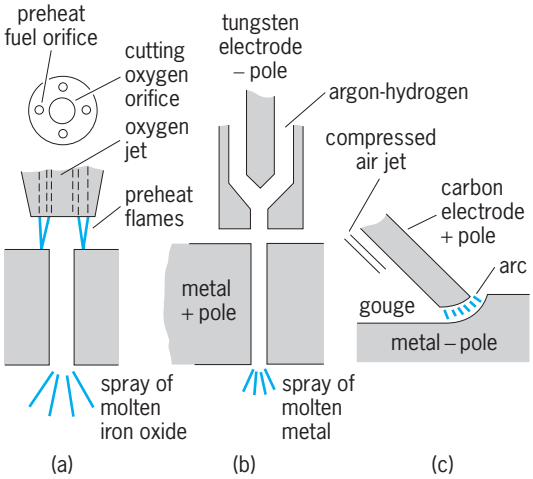


Fig. 14. Three thermal cutting processes. (*a*) Oxygen cutting. (*b*) Gas tungsten-arc cutting. (*c*) Air carbon-arc cutting.

Typical conditions for plasma arc cutting of aluminum

Thickness, in. (mm)	Speed, in./min (cm/min)	Orifice diameter, in. (mm)	Current, A	Gas flow,* ft^3/h (m^3/h)
1/4 (6.35)	300 (762)	1/8 (3.2)	300	80 (2.3) Ar–40 (1.1) H_2 130 (3.7) N_2–30 (0.85) H_2
1/2 (12.7)	200 (508)	1/8 (3.2)	300	65 (1.8) Ar–35 (0.99) H_2 140 (4.0) N_2–60 (1.7) H_2
1 (25.4)	90 (23)	5/32 (4.0)	400	65 (1.9) Ar–35 (0.99) H_2 140 (4.0) N_2–60 (1.7) H_2
2 (50.8)	20 (51)	5/32 (4.0)	400	65 (1.9) Ar–35 (0.99) H_2 140 (4.0) N_2–60 (1.7) H_2
3 (76)	15 (38)	3/16 (4.8)	450	130 (3.7) Ar–70 (2.0) H_2 140 (4.0) N_2–60 (1.7) H_2
4 (102)	12 (30)	3/16 (4.8)	450	130 (3.7) Ar–70 (2.0) H_2 140 (4.0) N_2–60 (1.7) H_2

*Either the mixture with argon-hydrogen or with nitrogen-hydrogen may be used.

except that an oxygen pipe replaces the torch for cutting steel up to 6 ft (1.8 m) thick. For highly alloyed and stainless steel cutting, it may be necessary to use a special process such as flux injection or powder cutting or some of the newer arc cutting processes described below. The selection of cutting process and type of operation (manual or mechanized) depends on the material that is being cut and the ultimate use of the product. In powder cutting, iron powder is injected into the oxygen in one process to increase the iron oxide content of the slag and thus lower its melting point.

There are seven cutting processes which utilize an arc. Some have been replaced by better and more efficient methods. These include oxygen arc cutting, air carbon-arc cutting, gas metal and gas tungsten-arc cutting, and plasma arc cutting.

In oxygen arc cutting, metals are severed by a combination of melting, oxidizing, diluting, fluxing, and mechanical ejection. For oxidation-resistant metals, the cutting mechanism is more of a melting action. Oxygen arc cutting electrodes were developed primarily for use in underwater cutting and were later applied for cutting in air. In both applications, they can cut ferrous and nonferrous metals in practically any thickness and position.

Air carbon-arc cutting is essentially a progressive melting of material which is blown away by high-velocity jets of compressed air. Base-metal melting is much less than in oxygen cutting. The process is primarily physical rather than chemical; hence it can be used on most metals.

Gas metal-arc cutting was developed soon after the commercial introduction of the gas metal-arc welding process. Conventional gas metal-arc welding equipment can be used for cutting, although some modifications have been made in the equipment designed especially for cutting operations. Gas tungsten-arc cutting with a nonconstricted arc can be used to sever nonferrous metals and stainless steel in thicknesses up to 1/2 in. (12.7 mm), using standard gas tungsten-arc welding equipment. Metals cut include aluminum, magnesium, copper, silicon bronze, nickel, copper nickel, and various stainless steels. This method can be used for either manual or mechanized cutting. Good-quality cuts can be made with an argon-hydrogen gas mixture containing 65–80% argon. Nitrogen can also be used, but the quality of cut is not as good as that produced with an argon-hydrogen mixture.

Plasma arc cutting employs an extremely high-temperature, high-velocity constricted arc between the electrode in the plasma torch and the piece to be cut. Where inert gases are used, the cutting process depends upon thermal action alone. When cutting such materials as mild steel and cast iron, increased cutting speeds can be achieved by using oxygen-bearing cutting gases. The process can be used to cut any metal (see **table**). The plasma cutting process can be used in the general fabrication of all metals in many industries, such as shipyard, aircraft, chemical, nuclear, and pressure-vessel. It also finds application in the forging and casting industry. The process can be used in manual or mechanized operation and is especially adaptable to automation.

The laser is a valuable tool for drilling, cutting, and eventually milling of virtually any metal or ceramic or other inorganic solid. The mechanism by which a laser beam removes material from the surface being worked usually involves a combination of melting and evaporation, although with some materials, such as carbon and certain ceramics, the mechanism is purely one of evaporation. Any solid material can be cut with the laser beam. In addition to the pulsed ruby laser, continuous CO_2 and pulsed YAG lasers are of importance in drilling, cutting, and trimming. The potential for laser cutting of reactive metals has been greatly enhanced by a technique where oxygen is added to the continuous-laser cutting process. On especially reactive material the cutting rate and penetration are greatly improved because of the exothermic reaction and creation of additional heat

due to the oxidation of the reactive material. *See* LASER.

Mel M. Schwartz

Bibliography. American Society for Metals, *Metals Handbook*, 10th ed., vol. 6: *Welding, Brazing, and Soldering*, 1993; American Welding Society, *Welding Handbook*, 8th ed., vol. 1, 1987, vol. 2, 1991, vol. 3, 1980; T. J. Blazynski, *Explosive Welding, Forming and Compaction*, 1983; H. Cary, *Modern Welding Technology*, 3d ed., 1993; H. Cerjak (ed.), *Mathematical Modeling of Weld Phenomena*, vols. 1–4, Institute of Materials, London, 1993, 1995, 1997, 1998; B. Crossland, *Explosive Welding of Metals and Its Application*, 1982; Deere and Co. Staff, *Welding*, 3d ed., 1991; D. S. Duvall, W. A. Owczarski, and D. F. Paulonis, TLP bonding: A new method for joining heat resistant alloys, *Welding J.*, 53(4):203–214, 1974; Ø. Grong, *Metallurgical Modelling of Welding*, 2d ed., Institute of Materials, London, 1997; T. B. Jefferson and D. T. Jefferson (eds.), *Welding Encyclopedia*, 18th ed., 1988; R. A. Lindberg and N. R. Braton, *Welding and Other Joining Processes*, 1985; G. R. Purdy (ed.), *Fundamentals and Applications of Ternary Diffusion*, Pergamon Press, New York, 1990; C. Stinchcomb, *Welding Technology Today: Principles and Practices*, 1989.

Well

An excavation made to extract water, oil, gas, brine, or other fluid substance from the earth. Wells are the source of about one quarter of the water supplies in the United States, all gas and oil, and most of the industrial brines and sulfur. Water wells are recorded in the earliest historic documents and probably originated during periods of drought when ancient humans attempted to reach water by digging a shallow excavation at the site of springs that had ceased to flow or in the dry channels of rivers. Later they learned to dig deeper for water where it did not issue at the land surface. The first wells were dug by hand, and some large deep ones were provided with elaborate ramps which enabled those drawing water to walk or even drive a donkey down to the water level. In the Near East for thousands of years wells (khanats) have been constructed by tunneling nearly horizontally into the outwash gravel to tap ground water which flows down the tunnel to its entrance. Horizontal wells are also used in Hawaii to skim fresh ground water from underlying salt ground water.

Most modern wells are of the drilled type. In the past the Chinese developed a crude percussion drill which enabled them to dig wells a few inches in diameter, lined with bamboo, to depths of a few thousand feet. In 1833 at Grenelle in Artois, France, a well was drilled to a depth of 1798 ft (548 m) and lined with steel casing. Its location provided the term artesian, which originally referred only to flowing wells. Drilling for brine in the United States began in West Virginia in 1808, and for oil and gas in Pennsylvania in 1859. Drilling for water in the United States began in the 1820s, but did not become common practice until the latter part of the nineteenth century. Dug wells are almost obsolete, because of the greater speed of drilling and the greater efficiency of drilled wells. *See* ARTESIAN SYSTEMS; OIL AND GAS WELL DRILLING.

Drilled wells, commonly 2–36 in. (5–90 cm) in diameter, usually are fitted with a steel tube or casing inserted in the drilled hole to the desired depth. Where the water-bearing formation is competent to stand without support, the casing is set, or finished, at the top of solid rock. Where there is danger of caving, as in sand or gravel, the casing is carried below the top of the water-bearing bed, and a perforated pipe or screen extends below the casing to the bottom of the bed. In fine-grained sand an envelope of coarse sand or fine gravel is commonly placed around the screen. In coarser-grained aquifers the screen is directly in contact with the aquifer. In either case the construction includes a considerable period of pumping, surging, or other treatment (called well development), during which the finer particles of the formation are drawn into the well and removed. This process substantially increases the initial yield of the well, in some cases by several times. Well screens may become encrusted with deposition of minerals from the ground water. Surging, acid treatment, or use of small explosive charges may be required to renew the well yield.

Most wells of large capacity are equipped with pumps of the deep-well turbine type to lift the water to the surface. However, where the water surface is less than 20 ft (6 m) below the land surface, they may be equipped with suction-type pumps. For efficiency the pump and motor should be selected to fit the expected discharge and lift. Many small-capacity farm and ranch wells are fitted with reciprocating pumps powered by windmills. Domestic wells are commonly equipped with electrically powered submersible centrifugal or jet pumps. If the well casing is bent or damaged, an air lift may be used. *See* PUMPING MACHINERY.

When a well is pumped, the pressure head at the well is lowered and a hydraulic gradient toward the well is established which causes water to flow toward the well. This lowering of head is called drawdown. One commonly used method for judging the ultimate yield of a well is to measure the yield per unit of drawdown; this is the specific capacity of the well, commonly expressed in gallons per minute per foot of drawdown as determined from pump tests after the well is completed. *See* GROUND-WATER HYDROLOGY.

Albert N. Sayre; Ray K. Linsley

Bibliography. J. Balek, *Groundwater Resources Assessment*, 1989; S. N. Davis and R. J. M. DeWiest, *Hydrogeology*, 1991; D. K. Todd, *Groundwater Hydrology*, 2d ed., 1980.

Well logging

The technique of making measurements in drill holes with probes designed to measure the physical and chemical properties of rocks and their contained fluids. Over a million holes are drilled annually around

TABLE 1. Logging methods

Logs and curves	Property investigated	Primary uses
Electrical		
Self-potential	Natural currents	Defining sand/shale sequences and bed thickness; determining correlation water salinity
Normal resistivity	Resistivity	Shallow investigation of invaded zone to aid in quantifying hydrocarbon aiding saturations; correlation
Lateral resistivity	Resistivity	Moderate to deep investigation of formation; used on older logging devices
Induction	Resistivity	Moderate to deep investigation of formations in high-resistivity muds
Focused	Resistivity	Shallow to deep investigation of formations in relatively low-resistivity muds
Microresistivity	Resistivity	Very shallow investigation to detect permeable rocks and aid in measurement of porosity
Dielectric constant	Dielectric properties	Distinguishing fresh water from oil in open hole
Microresistivity imaging	Resistivity	Obtaining two-dimensional microresistivity measurements giving gray scale or color pattern of borehole wall to produce corelike images; especially useful in identifying fractures, analyzing thinly bedded formations or structural features, and recognizing special features such as conglomerate pebbles and vugs
Acoustical		
Velocity	Interval transit time of compressional waves	Determining porosity, lithology, abnormal pressures; constructing synthetic seismograms; calibrating seismic records; generation of near-borehole seismic profiles
Wave train	Attenuation of acoustic waves	Detecting fractures, gas zones; determining cement bonding, mechanical properties of rocks
Nuclear		
Gamma ray	Natural radioactivity	Determining shale content, correlation, bed thickness; locating radioactive tracers
Spectral gamma ray	Energy spectrum of natural radioactivity	Identifying radioactive elements to help determine lithology
Density	Bulk density	Determining porosity; helping to identify gas zones
Neutron	Hydrogen content	Determining porosity; helping to identify gas zones
Pulsed neutron capture	Capture cross section of elements	Distinguishing between salt water and hydrocarbon behind casing
Geochemical spectrometry	Relative abundance of certain elements	Identifying clay types and other lithologies; detecting oil behind casing in presence of fresh water
Nuclear magnetism	Hydrogen atoms in larger pore spaces	Identifying productive zones; distinguishing heavy oil from fresh water
Production		
Flow meter	Fluid flow	Determining location and rate of fluid entry
Fluid density	Density of fluid in borehole	Determining type of fluid
Temperature	Temperature	Locating fluid entry, especially gas
Radioactive tracer	Fluid	Tracing flow path of radioactive material released in borehole
Noise	Fluid entry and flow	Detecting fluid movement inside or behind pipe and locating entry
Capacitance	Type fluid in borehole	Diagnosing fluid interfaces
Casing inspection	Electromagnetic, acoustic, or properties of casing	Evaluating corrosion, damage, or perforated holes in casing
Other		
Caliper	Borehole diameter	Estimating cement volumes behind casing; supplementing log analysis
Dipmeter	Azimuth and inclination of borehole and correlation of beds in borehole	Determining dip and strike of bedding, and geometry of hole
Borehole gravity meter	Bulk density	Determining density further from borehole; calibrating surface gravity surveys

the world to probe for oil, gas, solid minerals, and fresh water, and to better understand the subsurface geology. Much information can be obtained from samples of rock brought to the surface in cores or bit cuttings, or from other clues while drilling, such as penetration rate; but the greatest amount of information comes from well logs. *See* GEOPHYSICAL EXPLORATION.

Well logs result from a probe lowered into the borehole at the end of an insulated cable. While less common and generally more expensive, certain logs may also be recorded while drilling by means of special tools in the drill string above the bit (known as measurement while drilling, or MWD). Whether by cable or by measurement while drilling, the results are recorded graphically or digitally as a function of depth. These records are known as geophysical well logs, petrophysical logs, or more commonly well logs, or simply logs.

Recording information from boreholes has become more challenging as holes deviate radically from the vertical. It is becoming routine to log holes that are deviated 90° at a selected depth and penetrate over 2000 ft (610 m) in a horizontal direction.

An enormous number of logging devices have been developed since the first experiments with borehole resistivity measurements in 1927. From a few dozen basic logging tools, several hundred different types and variations of curves can now be recorded; however, in any one hole only about a dozen or fewer curves are commonly run. An outline of logging methods and some of their main applications are shown in **Table 1**.

In addition to these logging methods, there are a number of closely related electromechanical tools that can be run on the same wire line used for logging. Although these tools perform important functions in the borehole, they are not actually well logs. Some of these tools and their primary purposes are listed in **Table 2**.

Although the most common uses of logs are for correlation of geological strata and location of

TABLE 2. Electromechanical devices run on logging cables

Tool	Purpose
Sidewall sampling	Retrieving small core samples of formations from the side of an open borehole at selected intervals by using steel cylinders fired into the borehole wall (percussion samples) or a small rotary coring bit that penetrates the formation horizontally from a tool suspended in the well bore (rotary sample)
Wire-line tester	Testing formation pressures and fluids at selected intervals from cased hole or open hole
Perforators	Perforating casing for completion with bullets or shaped charge jets
Pipe cutters	Severing drill pipe or casing with mechanical, explosive, or chemical devices
Directional survey tool	Determining azimuth and inclination of borehole at selected intervals
Wire-line packer	Setting a plug in casing or tubing
Free-point indicator	Determining where pipe is stuck
Wireline shot	Loosening pipe so that it may be backed off threaded joint above where it is stuck
Fluid sampler	Catching a sample of fluid at a selected depth in borehole
Television or photographic cameras	Inspecting casing or other borehole conditions
Impression block	Determining nature of top of "fish" or lost tools in hole

hydrocarbon zones, there are many other important subsurface parameters that need to be detected or measured. Also, different borehole and formation conditions can require different tools to measure the same basic property. In petroleum engineering, logs are used to identify potential reservoir rock; determine bed thickness; determine porosity; estimate permeability; locate hydrocarbons; estimate water salinity; quantify amount of hydrocarbons; estimate type and rate of fluid production; estimate formation pressure; identify fracture zones; measure borehole inclination and azimuth; measure hole diameter; aid in setting casing; evaluate quality of cement bonding; locate entry, rate, and type of fluid into borehole; and trace material injected into formations (for example, to detect artificial fractures or to monitor injected fluids during supplementary recovery of hydrocarbons). In geology and geophysics, they are used to correlate between wells; locate faults; determine dip and strike of beds; identify lithology; deduce environmental deposition of sediments; determine thermal and pressure gradients; create synthetic seismograms; calibrate seismic amplitude anomalies to help identify hydrocarbons from surface geophysics; calibrate seismic with velocity surveys; generate vertical seismic profiles to give a more detailed seismic picture near the borehole; and calibrate gravity surveys with borehole gravity meter. Other applications include locating fresh-water aquifers; locating solid materials; and studying soil and rock conditions for foundations of large structures.

Electrical devices. These employ instruments that generate data based on electrical measurements. *See* ELECTRICAL MEASUREMENTS; ROCK, ELECTRICAL PROPERTIES OF.

Spontaneous potential. The spontaneous potential, usually recorded along with resistivity curves, is a simple but valuable aid to help geologists correlate from one well to another and to assist them in inferring the depositional environment of the sediments. It defines permeable zones from surrounding nonpermeable shales and can be used to estimate the salinity of formation water. In the early 1930s, observations were made in boreholes of naturally occurring currents. These currents, now referred to as spontaneous potential, self-potential, or simply SP, are recorded from a simple electrode on the probe with a reference electrode usually grounded at the surface. Spontaneous potential values opposite shales are relatively constant; however, when the electrode passes a permeable bed containing water more saline than the borehole fluid, the spontaneous potential will deflect to the left or in the negative direction (**Fig. 1**). The greater the salinity contrast between formation water and borehole water-based mud, the greater the spontaneous potential deflection. It is this electrochemical property that allows estimates of formation water salinity to be made.

Borehole size and bed thickness along with the density, chemical composition, and invasion of mud into the formation will also affect the magnitude of the spontaneous potential. The presence of hydrocarbons in the permeable zone may suppress the spontaneous potential a slight amount, but there is still formation water trapped in the smaller pores which will contribute to the spontaneous potential deflection.

Clays in permeable sand will also tend to suppress the spontaneous potential. Because the amount of clays deposited with sands can give valuable clues to the environment of deposition, the spontaneous potential shape is used by geologists to help piece together the subsurface geology. An example of the spontaneous potential curve, along with several other curves, is shown in Fig. 1. Note how the shaly sands tend to suppress the spontaneous potential toward the bottom of the lower sand interval. This sand could have been deposited as part of an ancient marine beach.

Resistivity. While resistivity was the first logging measurement to be made, it is still one of the most important physical properties to record. It is easy to visualize that a sand filled with salt water will conduct electricity much easier (and therefore have a low resistivity) than a sand with most of the salt water replaced with nonconducting oil or gas. The basic measurement of resistivity, the ohm-meter, is the electrical resistance of a cube 1 m on a side; it is the reciprocal of conductivity. Like most log measurements, the resistivity log response depends not only upon water saturation but also upon secondary properties of the rock/fluid system. Quantification

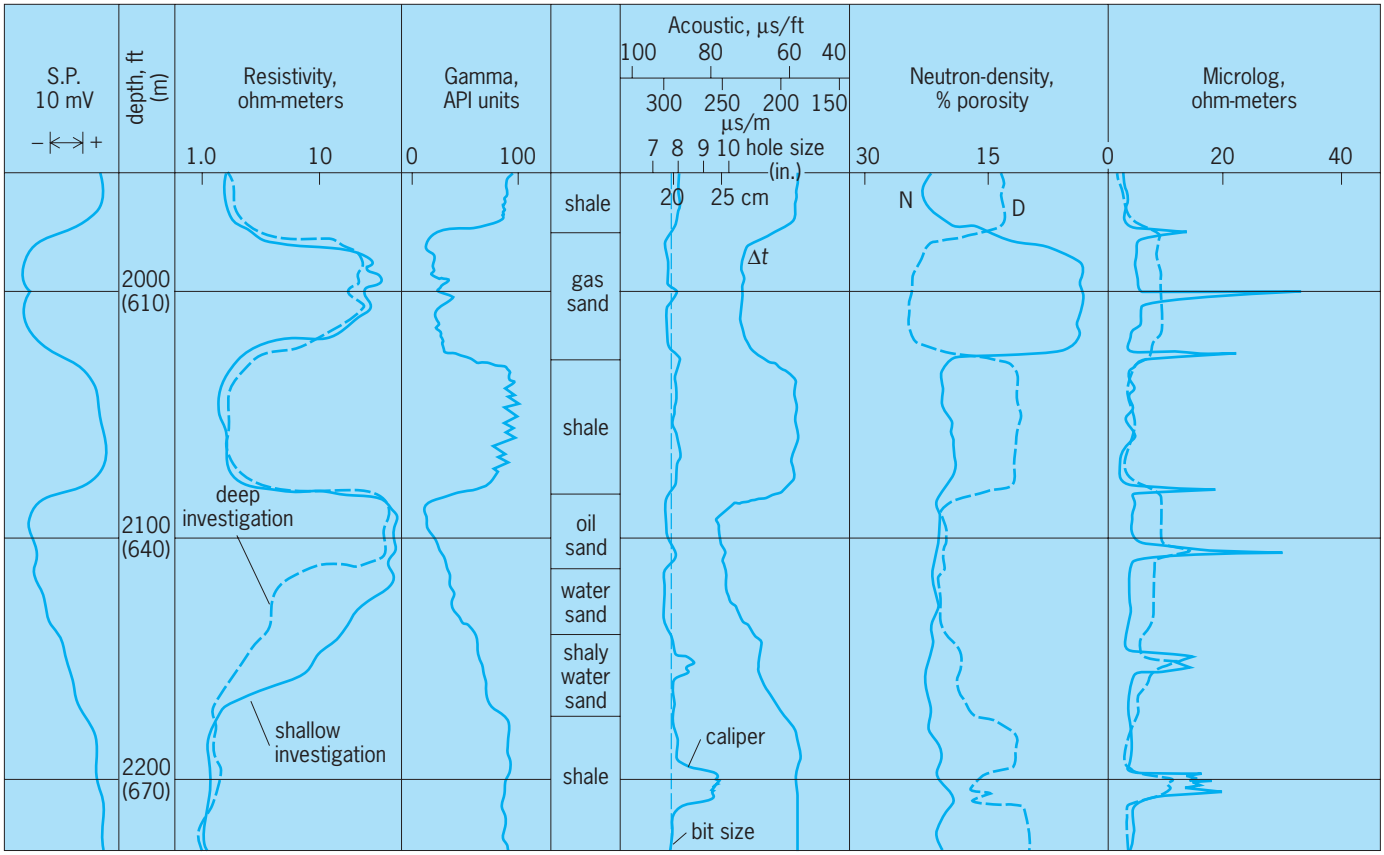


Fig. 1. Example of log response in sand/shale sequence.

of the water saturation S_w depends on knowledge of the rock pore space or porosity ϕ, cementation exponent m, saturation exponent n, and water resistivity R_w. In addition, the response characteristics of the particular tool for the surrounding physical environment, including mud and formation geometry, should be understood. In 1942 G. E. Archie published an empirical relation between water saturation, rock and fluid properties, and true resistivity R_t. This classic equation (1) is still widely used.

$$S_w = \left[\frac{R_w \phi^{-m}}{R_t}\right]^{1/n} \qquad (1)$$

Where significant amounts of conductive minerals such as clay are also in the formation, allowance must be made for their influence on the total resistivity. Many improvements have been made over the years in tools to record resistivity, but the primary purpose—to determine hydrocarbon saturation ($1 - S_w$) by calculation of the true formation resistivity—remains the same.

Some resistivity devices (listed below) are designed to investigate deep into the formation away from the borehole, while others respond to the formation near the borehole. In general, the resistivity tools with shallow investigation can define thin beds better, but are more influenced by the annular zone of mud filtrate invasion. The resistivity tools with deep investigation approach true resistivity, but still may need correction because of the influence of filtrate invasion, borehole geometry, and bed thickness.

1. Normal and lateral. The short normal, long normal, and lateral curves were common on older electrical logs and had shallow, intermediate, and deep investigations, respectively. The short normal, which has 16–18 in. (41–46 cm) spacing, is still used on the induction electrical log. In spite of its shallow investigation, it has been one of the basic geological correlation tools.

2. Induction log. The induction log is probably the most commonly run logging tool today; it is designed for medium and deep investigation. As shown in **Fig. 2**, this tool consists of one or more coils that generate an alternating electromagnetic field. The induced currents which are dependent upon the conductivity of the formation are then detected by receiver coils. Multiple coils are used to focus the measurement deep into the formation; this minimizes effects of the borehole, invaded zone, and nearby zones. The induction log works best in relatively resistive borehole fluids such as fresh-water or oil-based muds.

Modern induction tools are arrayed to enhance vertical resolution and radial investigation. Beds as thin as 1 ft can be resolved while providing images of formation resistivity that reflect bedding, hydrocarbon content, and invasion profiles. The display of this information can be enhanced be color images.

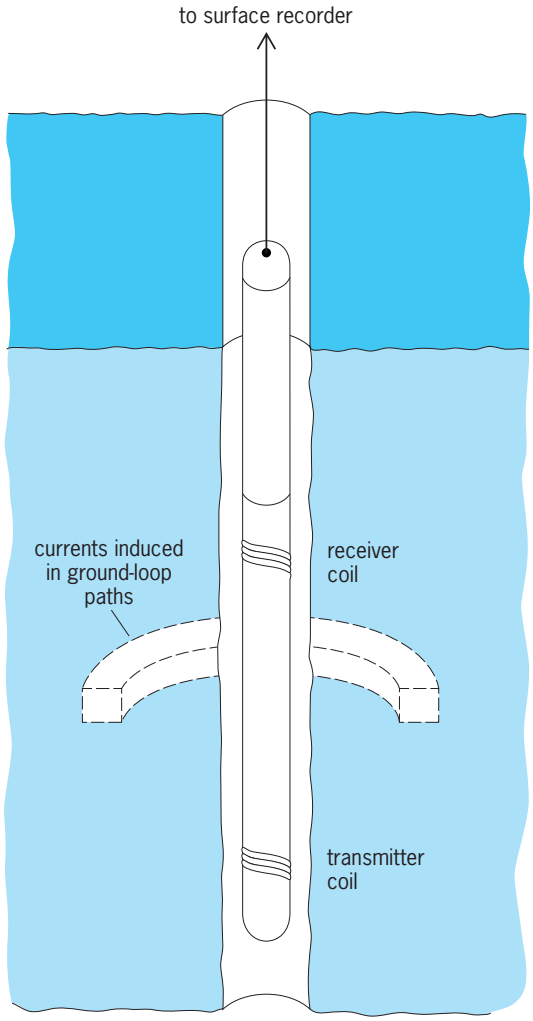


Fig. 2. Schematic of basic two-coil induction system.

3. Focused log. Where the borehole mud is more saline than the formation water, a focused log such as the Laterolog is recommended. In this tool a sheet of current from an electrode is forced into the formation by bucking currents from surrounding electrodes. This bucking current keeps the measuring current thin and allows excellent resolution of thin beds.

4. Microresistivity. There are several microresistivity tools where relatively shallow investigations of resistivity are made from pads pressed against the borehole wall. These auxiliary logs aid in the interpretation of true resistivity and porosity. One of these tools, the microlog (also known as minilog or contact log), is of particular interest. It is the oldest of all pad devices. This tool combines two resistivity measurements, one measuring about $1\frac{1}{2}$ in. (4 cm) away from the pad and the other about 4 in. (10 cm). When the pad is pressed against mud cake built up over a permeable zone, a separation appears between the two curves (Fig. 1). Shales or other nonpermeable zones do not have such separation; therefore, under favorable conditions the microlog can give excellent vertical resolution of permeable zones.

5. Microresistivity imaging. A tool to image the borehole wall electrically was introduced in 1985. This tool employs a dense array of sensors on a four-arm pad tool similar to the dipmeter. These sensors give a high-resolution electrical image of varying borehole wall conductivity that is displayed with a variable-intensity gray scale or in color. This gives an image of the borehole similar to a core photograph. The tool is especially useful in determining microheterogeneities such as fractures, bedding detail, pebbles, and vugs. This tool may give clues to productivity and depositional environment that, in absence of a core, could be missed by the other well-logging tools.

Acoustical devices. Quantitative interpretation of hydrocarbon saturation requires a knowledge of porosity. One of the most widely used porosity logs is the acoustic log, also known as a sonic log, acoustilog, or velocity log. This log is used for many other purposes, including identification of lithology, prediction of pressures, and assisting with geophysical interpretation. In addition, information from the amplitude of the acoustic wave aids in detection of fractures and gas zones, determination of mechanical properties of rocks, and analysis of cement bond behind pipe. *See* ACOUSTIC EMISSION.

The acoustical log measures the shortest time for sound waves to travel through 1 ft (or 1 m) of formation. This interval travel time [Δt; Eq (2)], measured

$$\Delta t = \frac{t_{(R_1)} + t_{(R_4)} - t_{(R_2)} - t_{(R_3)}}{4} \qquad (2)$$

in microseconds per foot (or per meter), is the inverse of velocity. A sketch of a borehole compen-

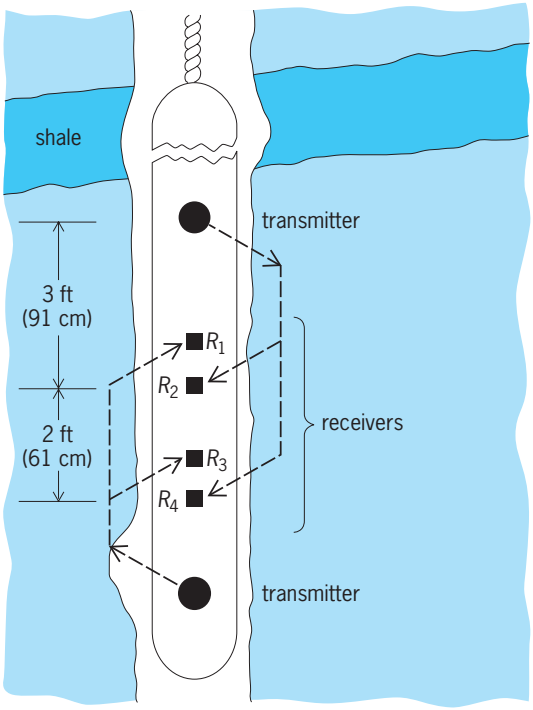


Fig. 3. Diagram of borehole compensated acoustic tool.

sated acoustic tool is shown in **Fig. 3**. The dual receivers with each transmitter are arranged so that travel time through mud and hole rugosity will not detract from measuring formation velocity. The formation travel time is dependent upon the porosity ϕ, the travel time of the fluids in the pore space (Δt_f), and the travel time of the rock matrix (Δt_{ma}). An empirical formula for clean sands known as the Wyllie time average equation relates porosity ϕ_{sv} to the sonic velocity interval travel time Δt of a water-filled rock as shown in Eq. (3). While Δt depends only

$$\phi_{sv} = \frac{\Delta t - \Delta t_{ma}}{\Delta t_f - \Delta t_{ma}} \tag{3}$$

upon the time of the first arrival of the compressional wave, other information may be extracted from the amplitude of the compressional wave and later arrivals.

In open hole logging, the amplitude of the acoustical wave train may be attenuated by natural fractures or faults. In cased hole logging, the cement bond log is an important aid in determining cement bonding in the annular space between the casing and the formation. This tool measures the amplitude of an acoustical signal imposed upon the casing. A good cement bond against the pipe will prevent it from ringing and therefore will greatly attenuate the amplitude of the compressional wave before it reaches the receiver. Supplemental information, including a display of acoustical wave trains, gives further insights into the degree of bonding.

Acoustical devices are closely linked to seismic interpretation. From the acoustical log (and density log if available) a synthetic seismic trace may be created to help correlate the seismic section with the detailed well-bore logs. By lowering multiple acoustical pickup devices or geophones into the borehole, while activating energy sources on the surface, acoustical travel time may be measured for seismic calibration. A similar arrangement is used for vertical seismic profiling, which generates a more detailed seismic section near the borehole. Vertical seismic profiling displays may also be generated from energy sources in the borehole, with geophones placed in the same borehole, at the surface, or in a nearby borehole (cross-well tomography). *See* SEISMIC EXPLORATION FOR OIL AND GAS.

Nuclear devices. These are instruments which generate data based on measurements of nuclear particles.

Gamma ray. In most cases the gamma-ray log may be considered as a shale log; that is, clays or shaly rocks will give a higher radioactive count than clean sands or carbonates. Like the spontaneous potential, this provides a good correlation curve, defines bed thicknesses, and aids interpretation of environmental deposition. It is usually a more diagnostic indicator of beds in a carbonate sequence than the spontaneous potential, and unlike spontaneous potential, can detect formations behind casing. The gamma log is relatively unaffected by borehole conditions. Often it is used to trace material injected into a formation, such as a radioactively tagged sand used as a proppant in a hydraulic fracture. Measurement is made in terms of arbitrary units, API (American Petroleum Institute) units. *See* GAMMA-RAY DETECTORS.

Spectral gamma ray. Almost all the naturally occurring radioactivity that is detected by the gamma-ray log comes from three elements: potassium, thorium, and uranium. Although it is usually assumed that high radioactive values on the gamma log come from potassium-rich shales, the spectral analysis log allows verification of the contribution from each element. Experience with this log has shown that a great number of "shales" are, in reality, uranium-rich stringers in sands, dolomites, and limestones. Some of these formations that were originally assumed to be impermeable shales have turned out to be hydrocarbon productive zones. The spectral breakdown also allows for more sophisticated geological environmental interpretation.

Density. Another important porosity tool is the density log; this tool emits a beam of gamma rays into the rock from a source such as cesium-137. These gamma rays interact with electrons in the formation through Compton scattering. The resulting lower-energy gamma rays are sensed by two detectors above the source. Although the tool responds to electron density, for most formation rocks the apparent bulk density is practically identical to the actual bulk density. For a few substances such as sylvite and rock salt (and to a lesser extent gypsum, anhydrite, and coal), small corrections are needed to arrive at true bulk density values. With knowledge of the bulk density ρ_b and reasonable assumptions of rock matrix density ρ_{ma} and fluid density ρ_f, the porosity derived from a density log (ϕ_D) can be calculated from Eq. (4).

$$\phi_D = \frac{\rho_{ma} - \rho_b}{\rho_{ma} - \rho_f} \tag{4}$$

An improved version of the density tool includes a measurement of photoelectric absorption. The magnitude of this measurement depends primarily upon lithology. However, if whole mud with barite, which has a very high photoelectric cross section, fills open natural fractures in a formation, the tool will respond accordingly. This effect, combined with high vertical resolution, can be used to identify open natural fractures.

Neutron. Another important device for determination of porosity is the neutron log. These logs respond primarily to hydrogen atoms. Therefore, in clean or shale-free formations the neutron log reflects the amount of liquid-filled porosity. Because the neutron log does not recognize gas-filled porosity, it can be compared with another porosity device such as the density to detect gas-filled zones. It can also be used in combination with other porosity logs for interpretation of lithology, including shaly sands. In neutron logging, high-energy neutrons are emitted from a source such as plutonium and beryllium. When these neutrons collide with hydrogen atoms which have a nucleus of nearly the same mass, the

greatest amount of energy is lost. When the neutrons have slowed down to thermal velocities after successive collisions, they are subject to capture by nuclei of other atoms; this causes an emission of gamma rays. Either these gamma rays or the low-energy neutrons are counted by the detector in the sands. *See* NEUTRON.

Pulsed neutron capture. Besides resistivity, another distinquishing physical characteristic is that the chlorine atom in salt water has a significantly higher capture cross section than the hydrogen or carbon atoms in oil or gas. Since resistivity measurements of the formation cannot be made in cased holes, a new tool sensitive to capture cross section was developed to distinguish hydrocarbon from salt water behind casing. The pulsed neutron capture tool employs a neutron generator that repeatedly emits pulses of high-energy neutrons. This is done by using a miniaccelerator to bombard a tritium target with deuterium atoms. After these neutrons are slowed down to the thermal state, they are captured by nuclei of the various atoms surrounding the tool. With each capture a corresponding emission of gamma rays occurs; these can be detected a short distance from the source. The total or bulk capture cross section (Σ_t) of the formation is the sum of the component cross sections of the rock matrix (Σ_{ma}) and the fluids such as water (Σ_w) and hydrocarbons (Σ_h) within the rock pores. This may be expressed as Eq. (5). Therefore, water

$$\Sigma_t = \Sigma_{ma}(1 - \phi) + \Sigma_w S_w \phi + \Sigma_b (1 - S_w)\phi \quad (5)$$

saturation S_w can be estimated where there is sufficient contrast in capture cross section between the water and hydrocarbons, and reasonable estimates can be made of the other variables. This tool has proved to be a valuable aid in identifying bypassed oil stringers in cased holes. Also it is used to monitor changes in the saltwater level within a producing horizon.

Geochemical spectrometry. There are various tools and combinations of tools designed to measure the abundance of elements commonly found in reservoir rocks. In general, these tools are designed to enhance but not replace the more traditional logging tools. Tools such as the carbon/oxygen log or gamma spectrometry tool can measure elements such as carbon, oxygen, calcium, silicon, hydrogen, iron, and sulfur and capture cross sections (sigma). In addition to giving more detailed lithologic information, these tools can directly measure hydrocarbons in zones with fresh or unknown-salinity waters in open hole or through casing. Since oil is composed largely of carbon, an abundance of this element in relation to others can provide a diagnostic indication of oil in the formation. This tool is similar to the pulsed neutron capture tool in that it utilizes a pulsed 14-MeV neutron source and gamma-ray detector. However, it is optimized for production of desired prompt gamma rays from inelastic scattering of neutrons by carbon in the formation. Because changes in lithology can also affect the carbon/oxygen ratio, a ratio of calcium/silicon is recorded to help distinguish a change in carbonate content from hydrocarbon saturation.

The geochemical logging tool combines many of the features of other tools, including the gamma spectrometry tool, the natural spectral gamma-ray tool, and the compensated neutron tool (which carries a special californium neutron source). In addition to the elemental analysis associated with such tools, a measurement can be made of aluminum weight concentration. Aluminum is a key element in clay minerals, micas, and feldspars. Titanium and gadolinium are also identified with this tool. It is possible to estimate concentrations of a missing element, for example, magnesium, not normally measured by spectrometric methods. Estimates of magnesium are important to identify dolomites. The wide variety of elemental data generated by a geochemical tool is valuable for detailed evaluation of a formation.

Nuclear magnetism log. The nuclear magnetism tool provides important auxiliary information for evaluation of subsurface fluids. Basically, measurements are made by applying a strong, magnetic field to the formation for a limited time (about 2 s). After the imposed field is relaxed, measurements are made of the free precession of protons (hydrogen atom nuclei) in pore space immediately adjacent to the borehole. Two independent parameters are derived from these measurements: free-fluid index I_{Ff} and thermal relaxation time T_1. Water molecules in very fine pores such as those found in clays will not respond. The magnitude of I_{Ff} is a value which approaches total porosity in clean formations with relatively large pores. The value of T_1 is dependent upon fluid–solid interactions and pore size. This tool is of great importance because of its ability to estimate permeability (at least within an order of magnitude). By killing signals from water treated with paramagnetic ions that have been injected into an oil zone, the nuclear magnetism log can measure the amount of residual oil with great accuracy.

Production logging. A very important suite of tools has been developed for production logging. These tools, usually run after the casing has been set, are designed to locate fluid movement behind the casing and into the well bore, to detect the type of fluid, and to determine the flow rate. Production logging tools include various types of flow meters, fluid density devices, sensitive thermometers, radioactive tracer devices, noise loggers, and capacitance logs. In addition, these tools usually record some correlation curve such as gamma-ray or collar locator to tie the production curves to the formation and casing. Closely related to production tools are logging devices that magnetically, acoustically, or mechanically inspect the condition of the casing in the borehole.

Other devices. There are a number of miscellaneous devices that give additional information.

Caliper. The caliper log is an important auxiliary curve that records the diameter of the hole with depth. This is essential for estimating the amount of cement needed in the annular space between the

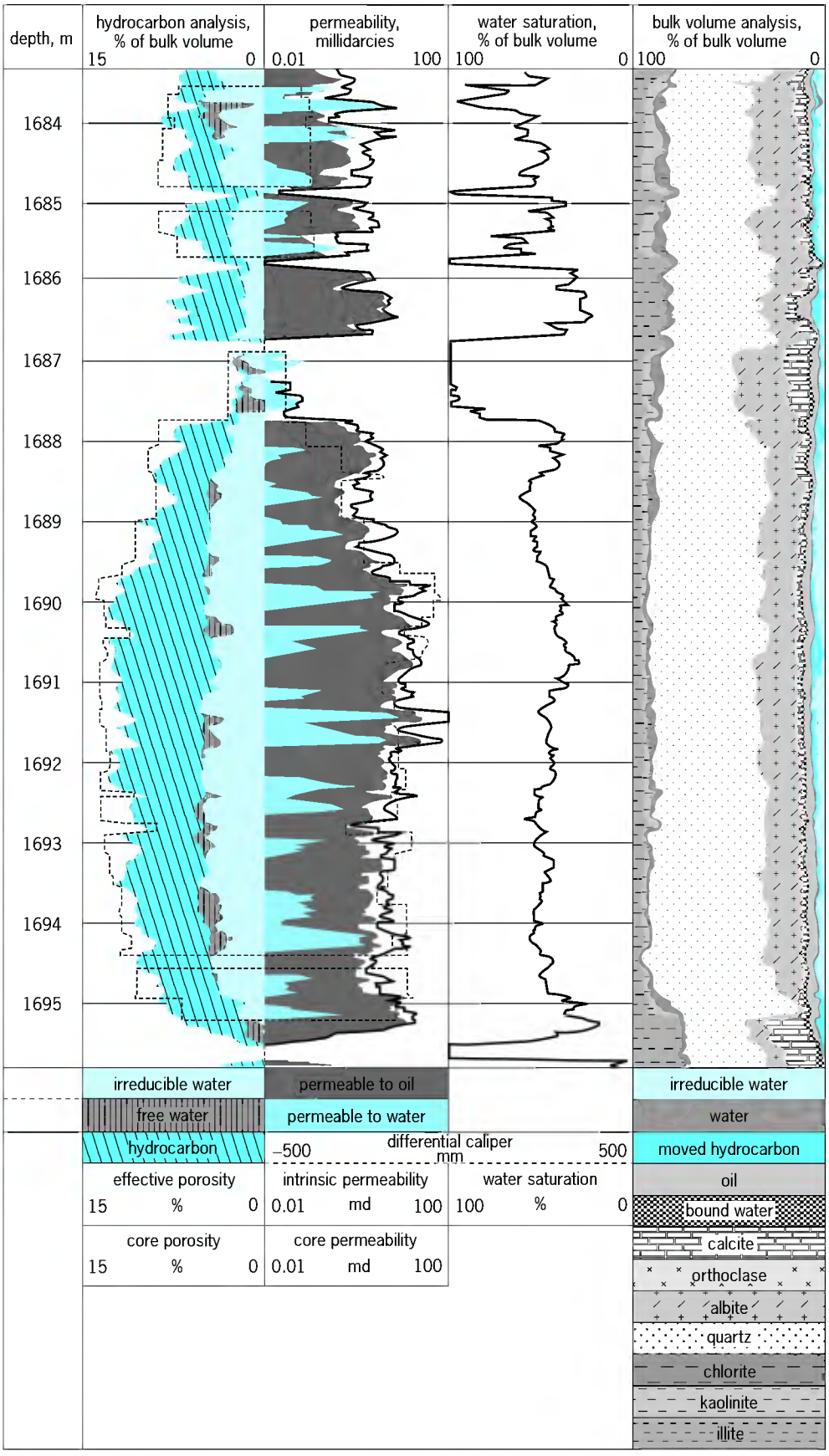


Fig. 4. Example of a computed log. In the hydrocarbon analysis column, the broken line represents core porosity. In the permeability column, the broken line represents core permeability, and the solid line represents intrinsic permeability. 1 m = 3.3 ft; 1 mm = 0.04 in. md = millidarcies.

casing and formation. It is also aids the interpretation of other curves where hole diameter or hole rugosity affects the readings. When a mud cake builds up over a permeable zone, the caliper can often detect such permeable streaks with excellent vertical resolution. The caliper is commonly run with each of the porosity tools.

Dipmeter. The dipmeter has three or four pads pressed against different sides of the hole. In the most common type of tool, these pads record microresistivity curves which can be correlated to determine depth displacement of formations from one side of the borehole to another. This information, combined with borehole size, deviation, and azimuth, is used to determine formation dip and azimuth. Such information is useful in correlating from well to well and in structural and stratigraphic interpretations. In addition, information necessary to plot the course of the borehole is provided by this tool.

Borehole gravity meter. Precise measurements of relative gravity can be made at intervals in the borehole to give estimates of formation density. These measurements are determined by large volumes of rocks that extend from tens to hundreds of feet away from the borehole. They are unaffected by borehole conditions such as mud type, hole rugosity, invaded zone, or casing and cement. Primary applications of the borehole gravity surveys include detection of significant porous zones or ore bodies away from the borehole, detection of oil and gas zones behind casing, and vertical density profiling to help interpret surface gravity mapping.

Dielectric constant. The dielectric constant log reflects the relative change in the dielectric constant of the rock fluids being surveyed. Since the relative dielectric constant of fresh water is significantly greater than oil, this tool can be used to distinguish oil from fresh water. The tools developed for open-hole logging measure phase and amplitude of a 16- and 30-MHz electromagnetic wave and at much higher frequencies such as 1.1×10^9 Hz. At these high frequencies, conductivity becomes much less important than dielectric permittivity, unlike the much lower frequencies used in the induction log. Because of the very shallow depth of investigation, this tool works best in smooth boreholes.

Computer analysis of logs. Well-log analysis has taken full advantage of the emergence of digital computing and the ability to take small, rugged, and powerful computers to the well site. Log data are recorded on magnetic disks or tapes as they are run, and measurements are provided through mathematical analysis. In addition, well-log data may be transmitted from the well site to the oil company headquarters through a high-speed digital satellite communication network. **Figure 4** illustrates an advanced petrophysical analysis with lithologic, fluid volume, and reservoir permeability calculations. Computation has been performed on a foot-by-foot basis with measurements from electric, nuclear, acoustic, and nuclear magnetism logs as well as rock cuttings from the borehole. Of particular note is the estimation of irreducible (nonmovable) water volume and log-derived permeability by interpretation of the nuclear magnetism and additional porosity logs. Rock-core measurements show good agreement with in-place log evaluation. In addition to analysis for open-hole logging, special programs are designed for cased-hole logging, solid mineral analysis, dip and structure of formations, synthetic seismic traces, and many others. In fact, the digital information from practically every curve can be incorporated into some type of analysis or interpretation program. *See* DIGITAL COMPUTER; MICROCOMPUTER.

Tools for measuring basic physical and chemical properties from a wellbore are now sophisticated and reliable. These tools combined with computer-aided modeling and interpretation programs have greatly enhanced the ability of oil and gas companies to evaluate subsurface geology, fluid content, and productivity.

Richard E. Wyman

Bibliography. T. Darling, *Well Logging and Formation EvaIuation (Gulf Drilling Guides)*, 2005; J. R. Hearst et al., *Well Logging for Physical Properties: A Handbook for Geophysicists, Geologists, and Engineers*, 2d ed., 2000; J. R. Jorden and F. L. Campbell, *Well Logging II: Electric and Acoustic Logging*, 1986; S. M. Luthi, *Geological Well Logs: Their Use in Reservoir Modeling*, 2001; R. Ransom, *Practical Formation Evaluation*, 1995.

Wellpoint systems

A method of keeping an excavated area dry by intercepting the flow of ground water with pipe wells located around the excavation area. Intercepting the flow before it reaches the excavated area also improves the stability of the edge of the excavation, permitting steeper bank slopes and often eliminating the need for supporting or shoring the banks. *See* CONSTRUCTION METHODS; GROUND-WATER HYDROLOGY; WELL.

Wellpoint systems are most effective in coarse-grained soils, such as gravel or sand. They are not effective in fine soils, such as silts and clays, where the small size of the pores between grains restricts the flow of water.

The basic components of a wellpoint system are the wellpoint, the riser pipe, the header pipe or manifold, and the pump (**Fig. 1**).

The wellpoint consists of a perforated pipe about 4 ft (1.2 m) long and about 2 in. (5 cm) in diameter. It is equipped with a ball valve to regulate the flow of water, a screen to prevent the entry of sand during pumping, and a jetting tip. The steel riser pipe to which the wellpoint is attached is slightly smaller in diameter than the wellpoint. Its purpose is to bring the ground water to the surface, where it is collected by the horizontal manifold pipe or header pipe, which is 6–10 in. (10–25 cm) in diameter. Wellpoints and riser pipes are spaced 3–6 ft (0.9–1.8 m) apart depending upon the characteristics of the ground. The wellpoint and riser pipe are

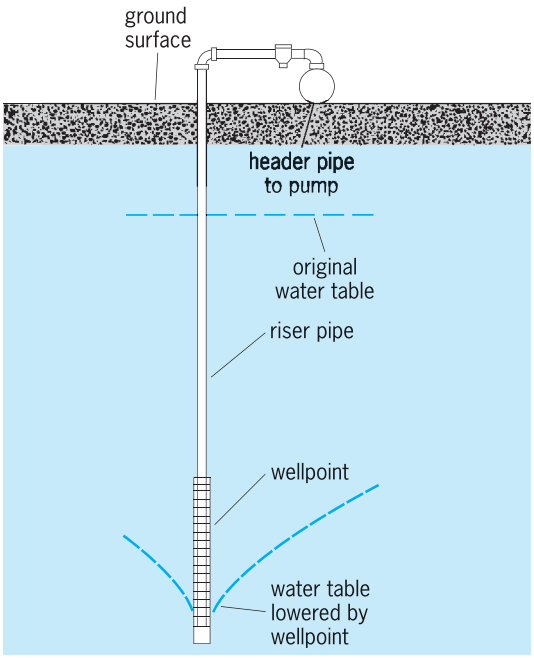


Fig. 1. Components of a wellpoint system.

Fig. 2. Layout of a three-tier wellpoint system. (*Griffin Wellpoint Co.*)

usually driven into the ground to the required depth, generally 2–5 ft (0.5–1.5 m) below the excavation, by jetting, that is, by displacing the ground below it with water under pressure.

Where the ground-water stratum is overlaid with hard clay or heavy gravel, jetting may not be able to pierce the tough material and a special boring attachment is used to carry the wellpoint down.

The pumps are located above the water table and collect the water from the header pipes for discharge away from the excavation area. The pumps are centrifugal pumps, with vacuum pumps to increase the suction lift. The suction lift in each wellpoint rarely exceeds 18–22 ft (5–7 m). Consequently, where the depth of the wellpoint below ground level exceeds this depth, it is necessary to install two or more tiers of wellpoint systems, each with its own assembly of wellpoints, header pipes, and pumps. A three-tier wellpoint system is shown in **Fig. 2**. *See* PUMP; PUMPING MACHINERY.

William Hershleder

Welwitschiales

An order of the class Cycadopsida having one species, *Welwitschia mirabilis*. This plant is native to the very arid deserts of southwestern Africa. Its appearance is bizarre—it has a very short, unbranched, woody stem (sometimes to 3 ft or 1 m in diameter), which is cushion- or saucer-shaped and tapers quickly to a long taproot. There are only two leaves, and these persist throughout the life of the plant. The leaf is broadly strap-shaped, as wide as the stem, firm, and leathery, and gradually splits lengthwise between the veins. The leaf (of indefinite growth) develops from a meristem at its point of connection to the stem. The species is dioecious; the cones are borne on branched axes originating between the crown of the stem and the base of the leaf. The order is known only from the one living species and not from any fossils. *See* CYCADOPSIDA; PINOPHYTA; PLANT KINGDOM.

Thomas A. Zanoni

Wentzel-Kramers-Brillouin method

A special technique for obtaining an approximation to the solutions of the one-dimensional time-independent Schrödinger equation, valid when the wavelength of the solution varies slowly with position; also known as the WKB method. It is named after G. Wentzel, H. A. Kramers, and L. Brillouin, who independently in 1962 contributed to its understanding in the quantum-mechanical application. It was however, studied earlier by J. Liouville (1837), Lord Rayleigh (1912), and H. Jeffreys (1923). It is also called the BWK method, and the JWKB method, the classical approximation, the quasiclassical approximation, and the phase integral method.

The system to be considered is a particle of mass m moving nonrelativistically in a potential $V(x)$. The Schrödinger equation for stationary states of energy E is Eq. (1), where $p^2(x)$ is defined by Eq. (2) and $\hbar$ is Planck's constant divided by 2π.

$$\frac{d^2\psi}{dx^2} + \frac{p^2(x)\psi}{\hbar^2} = 0 \tag{1}$$

$$p^2(x) = 2m[E - V(x)] \tag{2}$$

Two types of regions are to be considered as illustrated in **Fig. 1**: (I) The classically allowed type in which $p^2(x)$ is positive. Here $p(x)$ is defined to be the positive root, and it is the magnitude of the classical momentum. (II) The classically unallowed type in which $p^2(x)$ is negative. Here $p(x)$ is defined to be $i|p(x)|$. Typically $p^2(x)$ goes linearly through zero at the point where the region changes from classically allowed to unallowed. Such a point x_1, where $p^2(x_1) = 0$, is a turning point of the classical motion.

The solutions of differential equation (1) can be written, in a formal way, as Eq. (3). In this expression one uses all the upper signs uniformly or all the lower signs uniformly to obtain the two independent solutions. The constants in the indefinite integrals are the integration constants for the differential equation. The symbol $\sqrt{\ }$ indicates that the root of all that follows is to be taken, choosing the sign as suggested by the convention given above for $p(x)$. One verifies this formal expression of the solutions through Eqs. (4). Thus if the process indicated

$$\psi = \exp\left[\pm\frac{i}{\hbar}\int dx\sqrt{p^2(x)} \pm i\hbar\frac{d}{dx}\sqrt{p^2(x)} \pm i\hbar\frac{d}{dx}\sqrt{\cdots}\right] \quad (3)$$

$$\frac{d}{dx}\exp\left[\ \right] = \left[\pm\frac{i}{\hbar}\sqrt{p^2(x)} \pm i\hbar\frac{d}{dx}\sqrt{\cdots}\right]\exp\left[\ \right] \quad (4a)$$

$$\frac{d^2}{dx^2}\exp\left[\ \right] = \left[\pm\frac{i}{\hbar}\frac{d}{dx}\sqrt{p^2(x)\pm\cdots}\right]\exp\left[\ \right] + \left[-\frac{p^2(x)}{\hbar^2} \pm \frac{i}{\hbar}\frac{d}{dx}\sqrt{\cdots}\right] + \exp\left[\ \right] = -\frac{p^2(x)}{\hbar^2}\exp\left[\ \right] \quad (4b)$$

in Eq. (3) converges, it produces the solutions of the differential equation.

The WKB approximation to the solutions is found by evaluating the process to two orders only, Eq. (5). In a classically unallowed region the approximate solutions are usually written as Eq. (6).

$$\psi \simeq \exp\left[\pm\frac{i}{\hbar}\int dx\sqrt{p^2(x)\pm i\hbar\frac{dp}{dx}(x)}\right] \simeq \exp\left[\pm\frac{i}{\hbar}\int dx\left\{p\pm\frac{i\hbar}{2p}\frac{dp}{dx}\right\}\right] = \frac{1}{\sqrt{p}}\exp\left[\pm\frac{i}{\hbar}\int dx\,p(x)\right] \quad (5)$$

$$\psi = \frac{1}{\sqrt{|p|}}\exp\left[\mp\frac{1}{\hbar}\int dx|p(x)|\right] \quad (6)$$

Evidently the requirement for the applicability of the approximation is that $\hbar\,dp/dx$ should be small compared to p^2. In terms of the local de Broglie wavelength $\lambda = h/p$, this requirement is that $d\lambda/dx$ should be small compared to 2π. The WKB approximation therefore applies in the case of slowly varying wavelength.

Connection formulas. The approximation ordinarily applies in any region which does not include turning points, but it breaks down as $p^2(x)$ approaches zero, and in fact the approximate solution diverges as $p^{-1/2}$ at a turning point. There are connection formulas which relate the approximate solutions in a classically allowed region to the approximate solutions in a neighboring classically unallowed region. For the case illustrated in Fig. 1 the formulas are given by Eqs. (7). The integrals are always written in these

$$\frac{1}{\sqrt{p}}\sin\left[\frac{1}{\hbar}\int_x^{x_1} p(\xi)\,d\xi + \frac{\pi}{4}\right] \leftrightarrow \frac{1}{2}\frac{1}{\sqrt{|p|}}\exp\left[-\frac{1}{\hbar}\int_{x_1}^{x}|p(\xi)|\,d\xi\right] \quad (7a)$$

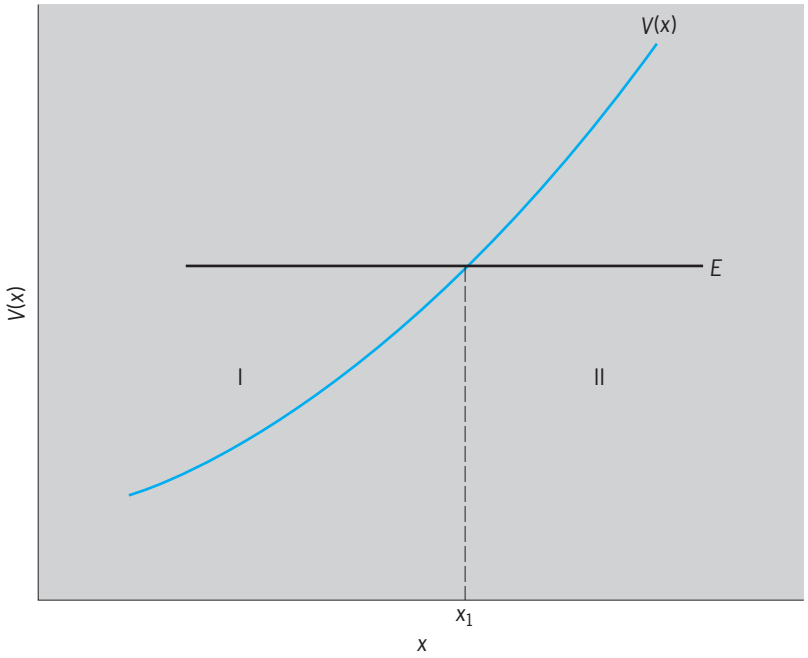


Fig. 1. Classically allowed region I to the left of a classically unallowed region II.

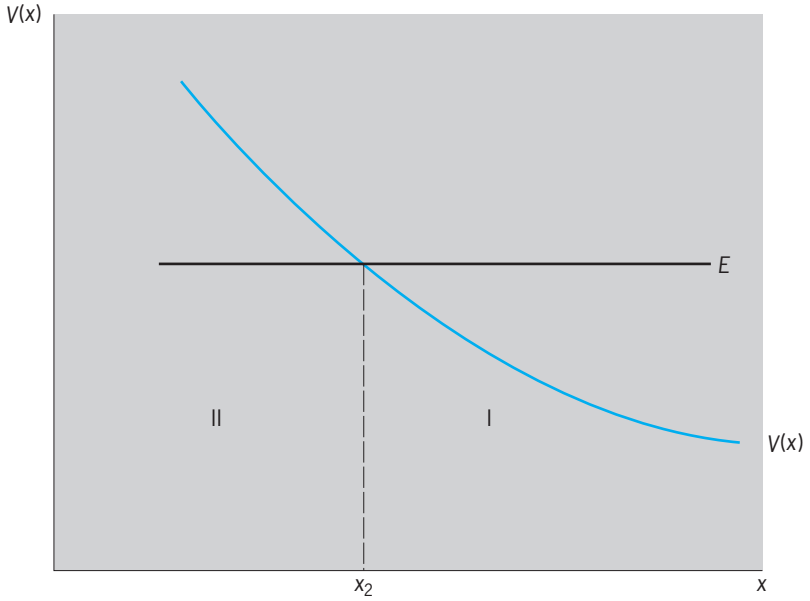


Fig. 2. Classically allowed region I to the right of classically unallowed region II.

$$\frac{1}{\sqrt{p}}\cos\left[\frac{1}{\hbar}\int_x^{x_1} p(\xi)\,d\xi + \frac{\pi}{4}\right] \leftrightarrow \frac{1}{\sqrt{|p|}}\exp\left[\frac{1}{\hbar}\int_x^{x_1} |p(\xi)|\,d\xi\right] \quad (7b)$$

formulas so that they are increasing functions of x away from the turning point. Similar formulas, given in Eqs. (8), apply when the allowed region is to the right, as illustrated in **Fig. 2**. The proof of these formulas is too long to be given in this article.

$$\frac{1}{2}\frac{1}{\sqrt{|p|}}\exp\left[-\frac{1}{\hbar}\int_x^{x_2} |p(\xi)|\,d\xi\right] \leftrightarrow \frac{1}{\sqrt{p}}\sin\left[\frac{1}{\hbar}\int_x^{x_2} |p(\xi)|\,d\xi + \frac{\pi}{4}\right] \quad (8a)$$

$$\frac{1}{\sqrt{|p|}}\exp\left[\frac{1}{\hbar}\int_x^{x_2} |p(\xi)|\,d\xi\right] \leftrightarrow \frac{1}{\sqrt{p}}\cos\left[\frac{1}{\hbar}\int_{x_2}^{x} p(\xi)\,d\xi + \frac{\pi}{4}\right] \quad (8b)$$

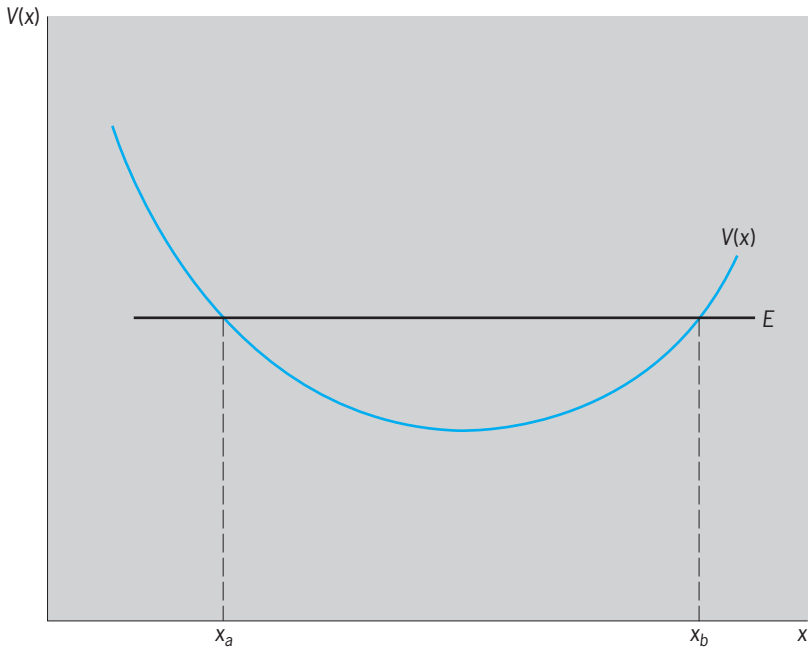


Fig. 3. Potential well with turning points at x_a and x_b.

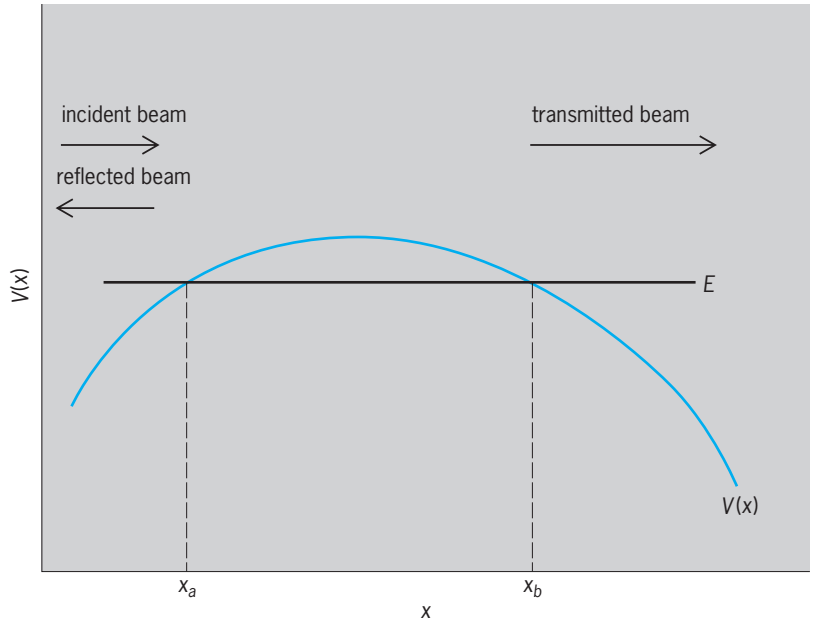


Fig. 4. Potential barrier with turning points at x_a and x_b.

Potential well. An important application of the WKB approximation is in obtaining an approximation to the allowed energy levels in a potential well. Suppose the well is as shown in **Fig. 3**. Only integrable functions, with decreasing exponential dependence, may occur in the classically unallowed regions to the left of x_a and to the right of x_b. Applying connection formulas (8*a*) and (7*a*) at the turning points x_a and x_b, one obtains two expressions for the approximate solution in the allowed region, Eq. (9). In order for these two expressions to coincide, it is necessary that Eq. (10) be satisfied, where

$$\frac{A}{\sqrt{p}}\sin\left[\frac{1}{\hbar}\int_{x_a}^{x} p(\xi)\,d\xi + \frac{\pi}{4}\right] = \frac{B}{\sqrt{p}}\sin\left[\frac{1}{\hbar}\int_x^{x_b} p(\xi)\,d\xi + \frac{\pi}{4}\right] \quad (9)$$

$$\frac{1}{\hbar}\int_{x_a}^{x_b} p(\xi)\,d\xi = \left\{n + \frac{1}{2}\right\}\pi \quad (10)$$

n indicates one of the values 0,1,2,3, ... This condition alternatively can be written as in Eq. (11). For

$$2\int_{x_a}^{x_b} \sqrt{2m\,[E - V(\xi)]}\,d\xi = \left\{n + \frac{1}{2}\right\}h \quad (11)$$

each allowed value of n there is an energy E_n that satisfies this condition; these are the allowed energy levels of the potential well, in the WKB approximation.

It is at this point that one can see the connection between modern quantum mechanics and the old quantum theory of N. Bohr and A. Sommerfeld. In the old quantum theory the allowed energy levels were found by setting the integral of the momentum, over a complete cycle, equal to an integer times Planck's constant. This condition coincides with Eq. (11), except that half-integer numbers are called for.

Barrier penetration. Another important application of the WKB approximation is in problems of barrier penetration. Consider a one-dimensional potential barrier as shown in **Fig. 4**, and suppose there is a particle beam of energy E incident on the barrier from the left. In the classical problem the particles of energy less than the peak potential energy would simply be reflected, but in the quantum-mechanical problem the particle beam is partly reflected and partly transmitted through the classically unallowed

region. The transmission coefficient D is defined as the ratio of the transmitted particle current to the incident current. By applying the connection formulas straightforwardly at the turning points and approximating for small D, one finds Eq. (12). This result

$$D = \exp\left[-\frac{2}{\hbar}\int_a^b |p(\xi)|\, d\xi\right] \qquad (12)$$

has wide applicability. It was used, for example, in the Folwer-Nordheim theory of field emission and in G. Gamow's explanation of alpha decay. *See* FIELD EMISSION; RADIOACTIVITY.

Generalization. The WKB approximation is based on the exponential function, the function that occurs in the exact solution of the free-particle Schrödinger equation. S. C. Miller and R. H. Good, Jr., developed a generalization of the WKB approximation, based on the functions that occur in the solution of an arbitrary comparison equation. The comparison equation is chosen so as to be similar to the equation under study and so as to be soluble, in the same spirit as perturbation theory. Qualitatively better approximations can be set up this way; for example, in the potential well problem, approximate solutions finite through both turning points can be found. *See* NONRELATIVISTIC QUANTUM THEORY; PERTURBATION (QUANTUM MECHANICS); QUANTUM MECHANICS. Roland H. Good, Jr.

Bibliography. V. P. Maslov, *The Complex WKB Method for Nonlinear Equations*, 1994; S. C. Miller and R. H. Good, Jr., A WKB-type approximation for the Schrödinger equation, *Phys. Rev.*, 91:174–179, 1953; E. L. Murphy and R. H. Good, Jr., WKB connection formulas, *J. Math. Phys.*, 43:251–254, 1964; D. Park, *Introduction to the Quantum Theory*, 3d ed., 1992; L. I. Schiff, *Quantum Mechanics*, 3d ed., 1968.

West Indies

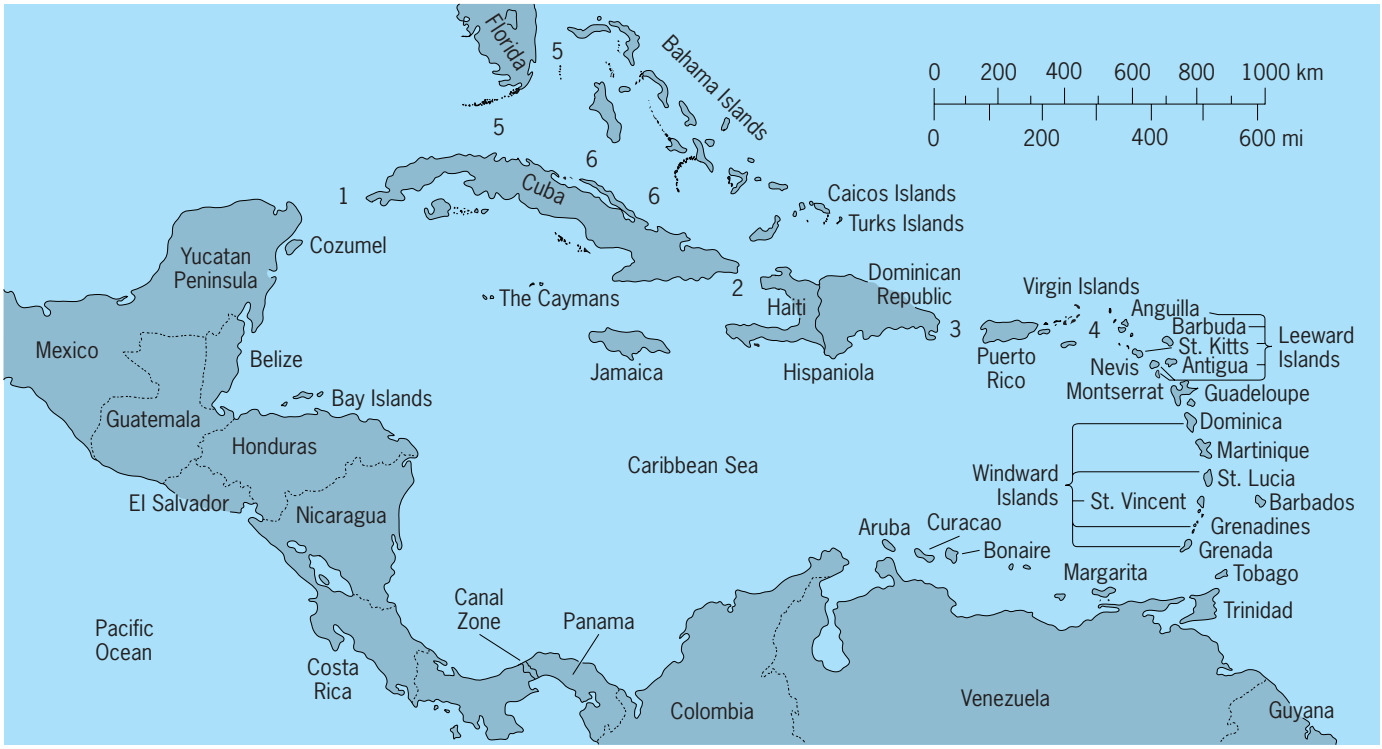
An archipelago, including the Bahamas, the Greater Antilles, the Lesser Antilles, and other islands, curving 2500 mi (4000 km) from Yucatan Peninsula and southeastern Florida to northern Venezuela and enclosing the Caribbean Sea (see **illus.**). Situated between latitude 10° and 27°N and longitude 59° and 85°W, in the zone of the northeast trade winds, the West Indies have a subtropical and predominantly oceanic climate, with even warmth and steady breezes. Temperatures vary little from season to season, ranging from means of 80–85°F (27–29°C) in July to 70–78°F (21–26°C) in January at sea level. Freezing is unknown, and the hottest temperatures rarely exceed 90°F (32°C). Precipitation ranges from a low of 25–50 in. (64–127 cm) a year on low-lying islands and drier coasts up to 300 in. (7.6 m) on the highest peaks, which are almost perpetually cloud-capped. At lower elevations, rainfall is erratic from year to year and from season to season, but reaches a maximum in the summer and fall, when the northeast trades are replaced by light, variable winds. This is also the season of hurricanes, destructive tropical cyclones which sweep west and northwest across the Caribbean, sparing only the southernmost islands. The winter months are generally dry, and there is frequently a shorter dry season in July or August. *See* TROPICAL METEOROLOGY.

The West Indian flora is chiefly derived from Central and South America, but there are a number of endemic species, notably palms; many mainland plants failed to colonize the islands, which are floristically poor. The effect of isolation and small size is evident in the meager character of West Indian fauna. Animal species are limited; there are few mammals and no large ones, except for domesticated animals and, especially on Hispaniola, feral cattle, goats, pigs, and horses.

The pre-Columbian flora and fauna of the islands, originally densely wooded, were largely transformed by European conquest and colonization. Sugarcane and cotton now occupy much of the lowlands, while bananas, tobacco, cacao, citrus, coffee, Old World pasture grasses, and American, African, and Asian subsistence crops mantle the foothills and mountain slopes. Mangroves and other salt-water vegetation still fringe most of the islands; cactus, thorn bush, and dry scrub occur on protected dry leeward sites, while other low-lying areas are savannas with scattered palms and pines. Higher up, where droughts are briefer and less intense, mixed conifer and broadleaf forests exist, with rainforest above. Windswept and sodden summits support palm brake and elfin woodland. *See* COTTON; MANGROVE; SUGARCANE.

Bahama Islands. The Bahamas (4404 mi^2 or 11,406 km^2, population 244,692), together with the Turks and Caicos islands (166 mi^2 or 430 km^2, population 2287) to the southwest, make up about 3000 islands, rocks, and cays stretching more than 600 mi (960 km) southward from Grand Bahama Island, 50 mi (80 km) east of West Palm Beach, Florida. Most of the islands (Andros Island, 1600 mi^2 or 4100 km^2 is largest) lie on the Great and Little Bahama banks, slightly submerged limestone platforms separated from Florida and Cuba by the shallow Straits of Florida and Old Bahama Channel. Made of calcareous sand of marine origin and flanked on the east by a coral barrier reef, the Bahamas are flat and low, exceeding 200 ft (60 m) in height only on Cat Island (400 ft or 120 m). Precipitation is consequently scanty and unreliable, ranging from 40 in./year (102 cm/year) at Nassau, New Providence Island, to 25 in. (64 cm) at Grand Turk. What little soil exists is generally fertile, but the significant produce of land and sea is salt and sponges. Formerly a British colony, the Bahamas became independent in 1973; the Turks and Caicos islands are British.

Greater Antilles. These islands constitute the major share of the West Indies. They represent a partly submerged extension of an ancient Central American mountain chain, which stretches eastward from Hispaniola beneath 75-mi-wide (120-km) Mona Passage through Puerto Rico and the Virgin Islands to 40-mi-wide (64-km) Anegada Passage. Westward from Hispaniola the range splits in two. The southern



Island locations of the West Indies. Outstanding inter-island seaways are 1, Yucatan Channel; 2, Windward Passage; 3, Mona Passage; 4, Anegada Passage; 5, Straits of Florida; and 6, Old Bahama Channel. 1 mi = 1.6 km.

prong runs from the southwestern peninsula of Haiti into the sea, emerges in the Blue Mountains of Jamaica, and reappears to the west in Swan and Roatán islands off northeast Honduras. The northern prong runs from Haiti's northwest peninsula across 55-mi-wide (89-km) Windward Passage, reappearing in the Sierra Maestra of southeastern Cuba and again, further west, in the Cayman Islands (100 mi^2 or 259 km^2; population 10,419, British), low, flat limestone islands with fringing coral reefs, and reaches the shore of Central America north of the Bay of Honduras. Between these two prongs runs Cayman Trough (Bartlett Deep and Oriente Deep), which reaches 23,748 ft (7238 m) below sea level. North of the entire structure, separating Puerto Rico from the Bahama platform, is Puerto Rico Trench, 30,180 ft (9199 m) below sea level in Milkwaukee Depth. The extremes of elevation, the intense faulting and folding of most of the strata, gravity anomalies, and earthquakes all attest to the tremendous pressures to which the whole area has been subjected.

Cuba. With a length about 760 mi (1220 km) and area of 44,218 mi^2 (114,524 km^2; population 10,073,000), this is the largest of the West Indies. Only its southeastern mountain range, the Sierra Maestra, which reaches 6429 ft (1960 m) at Pico Turquino, links it structurally and physiographically with the other Greater Antilles. The rest of the long, narrow island is essentially a level or gently sloping limestone plateau resembling southern Florida, the Bahamas, and, west of the broad but shallow Yucatan Channel, the peninsula of Yucatan. The low-lying plateau is significantly broken by the Grupo de Guamuhayo (3792 ft or 1156 m at Pico San Juan) in the center of the island and by the Cordillera de los Organos at the western end, the latter a broken limestone country of karst topography, with landforms produced by chemical solution rather than by subaerial erosion. Lacking a central dividing mountain range, Cuba gets less rainfall (about 50 in. or 125 cm annually) than others of the Greater Antilles, but since there is no impediment to the trade winds, the south coast receives as much as the north. Only the western end—closest to the usual hurricane track—is much wetter. The typical vegetation is a thorn-scrub steppe of pine and palmetto interspersed with short grass. Among numerous offshore islands, the Isle of Pines (862 mi^2 or 2233 km^2), to the southwest, is largest. *See* KARST TOPOGRAPHY.

Jamaica. This island (4244 mi^2 or 10,991 km^2; population 2,214,120) is built up on an east-west trending igneous and metamorphic core, which reaches its greatest elevation in the Blue Mountains (Blue Mountain Peak, 7402 ft or 2256 m) at the eastern end of the island and is exposed at lower heights in the center. These steep and rugged mountains are subject to landslides and severe erosion where deforested. To the west and on their north and south flanks they are overlain, at about 3000 ft (900 m) above sea level, by a limestone plateau. But the limestone is so intricately dissected that little remains of the plateau surface. Karst topography is widespread and in the Cockpit Country of north-central Jamaica takes the form of sinkholes 500 ft (150 m) deep and a quarter mile (400 m) across. A by-product of weathering in some of the limestone glades is bauxite ore;

Jamaica is estimated to possess 600,000,000 commercially utilizable tons (550,000,000 metric tons). Thanks to their high alumina content, their surface occurrence, and their proximity to North American reduction plants, the Jamaican deposits are among the world's most productive. Level land is practically confined to alluvial plains, of which the largest is along the south-central coast. Formerly a British colony, Jamaica became an independent nation in 1962.

Hispaniola. With a total area of 29,530 mi^2 (76,484 km^2), the eastern two-thirds of this island is the Dominican Republic (population 5,518,430); the western third is Haiti (population 5,054,090). It is the most mountainous and structurally complex of the Greater Antilles. The Cordillera Central traverses the island from southeast to northwest at 8000–9000 ft (2400–2700 m), culminating in Pico Trujillo (about 10,200 ft or 3100 m), the highest elevation in the West Indies. There are two other major ranges. In the north, entirely within the Dominican Republic, the 4000-ft (1200-m) Cordillera Septentrionale runs eastward into the Peninsula of Samaná. In the southwest, another range forms the Tiburón Peninsula of Haiti and extends eastward into the Sierra Baharuco of the Dominican Republic; though narrower, it is almost as high as the Cordillera Central, reaching 7920 ft (2414 m) in the western Massif de la Hotte and 8793 ft (2680 m) toward the east in the Massif de la Selle. Between the southwest and central ranges runs a deep structural trench, the Cul de Sac, a continuation of the Cayman Trough between Cuba and Jamaica. The Cul de Sac contains two brackish lakes, of which the easternmost—Lake Enriquillo in the Dominican Republic—is 160 ft (49 m) below sea level.

Besides the Cul de Sac, the only extensive level areas in Hispaniola are the plain separating the northern and central ranges (Plaine du Nord in Haiti, the Cibão in the Dominican Republic), the extensive Seibo lowland in southeastern Dominican Republic, and the plain of the Artibonite River at the western end of the Cordillera Central. Above the last at about 1000 ft (300 m) lies a rolling intermontane basin, the Plaine Centrale. The largest offshore island, Ile de la Gonave (254 mi^2 or 658 km^2, population 27,000), lies in the Golfe des Gonaïves, almost encircled by the two western peninsulas; Ile de la Tortue (70 mi^2 or 181 km^2) lies off Cap Haitien to the north.

Hispaniola's climate is less oceanic than that of the other West Indies. The trade winds fail to penetrate to many sheltered locations, which warm up like continental areas in summer. Because of the heat, some of the plains are subject to drought even though they receive considerable rainfall; the Cul de Sac gets almost 50 in. or 127 cm annually, but evaporation is so intense that the vegetation is of semiarid type.

Puerto Rico. An intricately dissected east-west trending mountain core dominates the topography of this island (3423 mi^2 or 8866 km^2; population 3,187,550). The highest part rises to 4400 ft (1341 m) in Cerro de Punta. To the south is a narrow, arid coastal plain. The northern coastal plain, broader and comparatively well watered, is separated from the mountains by dissected limestone terraces, except in the northeast, where the Sierra de Luquillo (3500 ft or 1067 m), an outlier of the central range, reaches close to the sea. Little of the natural vegetation remains anywhere. A Spanish dependency until 1898, Puerto Rico, together with its off-shore islands, Mono (20 mi^2 or 52 km^2) to the west, Vieques (51 mi^2 or 132 km^2; population 10,000), and Culebra (11 mi^2 or 28 km^2; population 1000) to the east, has since been an American possession and is now a self-governing commonwealth under the United States.

Virgin Islands. These small islands with a total area of about 200 mi^2 (518 km^2) appear as eastern outliers of the Puerto Rico mountain mass but were submerged much more recently and hence display marine physical features—notable barrier reefs—superimposed on continental rocks. Anegada, at the extreme northeast of the Greater Antillean chain, is a flat coral reef raised 15 ft (4.5 m) above sea level; the other islands have steep slopes and a minimum of flat land, except for St. Croix, south of the main range. Maximum elevation, however, is only 1780 ft (543 m), at Mount Sage (Tortola); rainfall is low and erratic; vegetation consists of a scanty dry scrub woodland. The Virgin Islands are divided between the United States (St. Croix, St. Thomas, and St. John, total 133 mi^2 or 344 km^2; population 71,236) and Great Britain (Tortola, Virgin Gorda, Jost Van Dyke, and about 50 other islets of which six are inhabited, in all 59 mi^2 or 153 km^2; population 14,024).

Lesser Antilles. The numerous islands with a total area of about 2500 mi^2 or 6500 km^2 (population 1,406,000) compose an arc marking the eastern margins of the Caribbean Sea. Built up on a submarine ridge, they stretch 500 mi (800 km) from the eastern end of the Greater Antilles at Anegada Passage in the north to a 90-mi-wide (145-km) channel between Grenada and Tobago in the south. Mostly oceanic, the Lesser Antilles are festooned in two, perhaps three, separate chains: an inner ring of volcanic islands; an outer line of predominantly limestone islands to the east; and Barbados, farthest east. The largest, Guadeloupe, is a composite: the western half (Basse Terre) volcanic, the eastern (Grande Terre) limestone. From north to south, the main islands in the two chains are shown in the **table**. Formerly British colonies, Barbados became independent in 1966, Grenada in 1974, Dominica in 1978, and St. Lucia and St. Vincent (including the Grenadines) in 1979; except Montserrat, the remaining British islands are self-governing Associated States. Martinique and Guadeloupe, with the latter's small dependencies, are departments of France. The Dutch islands belong to the Netherlands Antilles.

Main volcanic arc. In this chain, islands are spaced at regular intervals, with channels between them nowhere more than 30 mi (48 km) wide. With hard basaltic cores covered by ash and other igneous material, these islands are the remnants of volcanoes in various stages of decay and subaerial erosion. At the north and south ends of the chain there has

Lesser Antilles

Islands	Maximum elevation, ft (m)	Area, mi^2 (km^2)
Main volcanic arc		
Saba (Dutch)	2850 (869)	5 (13)
St. Eustatius (Dutch)	1980 (604)	8 (21)
St. Kitts (Br.)	3711 (1131)	65 (168)
Nevis (Br.)	3300 (1006)	36 (93)
Redonda (Br.)	975 (297)	0.3 (0.8)
Montserrat (Br.)	3002 (915)	39 (101)
Guadeloupe, west (Fr.)	4869 (1484)	327 (847)
Les Saintes (Fr.)	1036 (316)	5 (13)
Dominica	4747 (1447)	290 (751)
Martinique (Fr.)	4429 (1350)	417 (1080)
St. Lucia	3145 (959)	238 (616)
St. Vincent	4048 (1234)	133 (344)
The Grenadines	980 (299)	30 (78)
Grenada	2750 (838)	119 (308)
Outliers		
Sombrero (Br.)	40 (12)	0.2 (0.5)
Anguilla (Br.)	200 (61)	35 (91)
St. Martin (half Fr., half Dutch)	1350 (411)	33 (85)
St. Barthélemy (Fr.)	990 (302)	8 (21)
Barbuda (Br.)	200 (61)	60 (155)
Antigua (Br.)	1319 (402)	108 (280)
Guadeloupe, east (Fr.)	1300 (396)	226 (585)
La Désirade (Fr.)	930 (283)	8 (21)
Marie Galante (Fr.)	670 (204)	61 (158)
Barbados	1104 (336)	167 (433)

been no recorded post-Columbian volcanic activity, but from St. Kitts to the Grenadines the cones are still active; in 1902, St. Vincent's Soufrière and Martinique's Mount Pelée erupted, causing widespread damage and killing 30,000 Martiniquans. Smaller eruptions occurred in Guadeloupe (1956) and north of Grenada (1966 submarine). Solfataras, or emissions of hot sulfurous vapors and gases, and fumaroles, or eruptions of steam from subsurface lava, evince continued volcanic activity, and earthquake shocks are frequent. All extremely mountainous, these islands rise in some cases thousands of feet sheer out of the ocean; the Pitons of St. Lucia are the most spectacular example. There is little level land, and the dense populations (400/mi^2 or 150/km^2) occupy all but a small part of even the steepest slopes.

Outliers. By way of contrast, the outlying islands are low-lying and flat or only moderately hilly. They consist of ancient submerged volcanoes capped by sedimentary strata, now slightly tilted. Less fertile and more arid than the inner volcanic arc, they support only a degraded scrub vegetation and are, for the most part, less densely populated than the islands to the west.

Barbados. Located at the extreme southeast of the chain of outliers, this is remnant of an ancient continent. Coral limestone escarpments ascend to an elevation of 1104 ft (336 m) in Mount Hillaby. Sedimentary strata of continental origin underlie the coral and appear at the surface in the rugged, deeply dissected Scotland District of the northeast. Elsewhere, the undulating and intensely cultivated surface is reminiscent of an English landscape.

Offshore islands. There are a number of islands near the coasts of Central and South America.

Trinidad and Tobago. With a combined area of 1981 mi^2 (5130 km^2) and population of 1,179,100, these are detached fragments of the Caribbean Coastal Range of the Andes in northern Venezuela, from which Trinidad is separated by the Gulf of Paria. In climate and culture they resemble the Lesser Antilles, but their landforms and vegetation are more continental than insular. They gained independence from the British in 1962.

Trinidad (1864 mi^2 or 4828 km^2) is made up of three mountain ranges with lowlands and swamps between them. The forested northern range, which reaches a height of 3085 ft (940 m), is an east-west trending continuation of the Península de Paria of Venezuela to the west, cut off by the Bocas del Dragón and a chain of small islands. The central range bisects the island from southeast to northwest; the southern range, like the central range only 1000 ft (300 m) high, forms the southern rim of Trinidad, separated from Orinoco delta by the Boca de la Sierpe. Between the central and southern ranges are deposits of asphalt, in Pitch Lake, and of oil, with the refining of imported oil the mainstay of the local economy.

Tobago (116 mi^2 or 300 km^2) is aligned along a densely wooded ridge which rises to 1800 ft (550 m) in the northeast and slopes off to the southwest.

Southern Caribbean. Off northern South America several islands appear, like Trinidad and Tobago, as outliers of Andean ranges. Venezuelan Margarita (444 mi^2 or 1150 km^2), a mountainous island which reaches 4800 ft (1460 m) at its western end, and flat Tortuga are remnants of the Caribbean Coastal Range. Farther west, the Netherlands Leeward Islands of Aruba, Bonaire, and Curaçao (71, 120, and 178 mi^2, or 184, 311, and 461 km^2), together with a few small Venezuelan islands, represent an extension of the peninsula of Guajira in Colombia. The Dutch islands are worn-down remnants of crystalline rocks, from which most of an overlying limestone has been eroded. Low-lying and exposed to the desiccating sweep of the northeast trades, they are arid (mean annual rainfall 20 in. or 50 cm) and support only xerophytic, or drought-resistant, vegetation.

Western Caribbean. At the western end of the Caribbean lie San Andrés (9 mi^2 or 23 km^2) and Providencia (12 mi^2 or 31 km^2) east of Nicaragua but dependencies of Colombia; and the Swan Islands (2 mi^2 or 5 km^2), northwest of Nicaragua but United States possessions. Much closer to the Central American coast are the San Blas, or Mulatas, Islands of Panama; the Corn Islands (6 mi^2 or 16 km^2) of Nicaragua, on lease to the United States; the Bay Islands of Honduras and British Honduras; and Cozumel (189 mi^2 or 490 mi^2 including its outlier Mujeres) of Mexico, east of Yucatan. Providencia is a complex volcanic plug (1200 ft or 365 m), San Andrés and Swan are limestone similar to the Caymans, the Corn Islands are of porous lava, and the Bay Islands are outliers of the Central American range that emerges farther east in Jamaica.

David Lowenthal

West Nile Virus

An arbovirus, first identified in the West Nile area of Uganda in the early 1930s, that has been an increasing threat in North America since the late 1990s. Arboviruses (***ar***thropod-***bo***rne viruses) are carried by arthropods and transmitted to the host through the bite of the insect. West Nile virus (WNV) found its way to North America in 1999, possibly through international travel, the importation of infected birds or mosquitoes, or migration of infected birds. Severe infection with WNV can result in viral encephalitis, a dangerous and sometimes fatal inflammation of the brain. Since 1999, WNV infections in humans, birds, and mosquitoes have been reported from all states except Hawaii, Alaska, and Oregon. In 2004, there were 2539 cases of human WNV infection in 41 states in the United States. The incidence of WNV encephalitis in the United States is seasonal, with peaks corresponding to the times of the year during which adult mosquitoes are active.

Transmission and epidemiology. In order to be effectively transmitted, arboviruses must have the ability to infect vertebrates (such as birds or humans) as well as the invertebrate vector. The virus must then be able to grow in the vertebrate's bloodstream and multiply (this is known as a "viremia"), remaining in the blood long enough for infection of a second invertebrate vector. Finally, it must infect the second invertebrate's salivary gland to allow transmission to other vertebrate hosts. For WNV, this cycle occurs when the *Culex* mosquito bites an infected bird, taking up the virus and becoming infected in its turn, after which the mosquito bites another bird and continues the transmission cycle.

Humans are usually "dead-end hosts," meaning they cannot spread WNV back to the vector because they do not maintain a persistent viremia. It is now recognized, however, that the virus can be transmitted by blood transfusions and organ transplants. Despite this development, the reservoirs for this virus are birds, not humans. Reservoirs are susceptible hosts that allow the reinfection of other arthropods. Therefore, there is no transmission of WNV from human to mosquito. WNV has been detected in at least 138 species of birds. Although birds infected with WNV may become ill or die (especially crows and jays), most survive infection. Bird reservoirs that are hosts for the virus will experience an infectious viremia for 1 to 4 days after exposure. After this time, the bird will either succumb to the infection or will develop life-long immunity to WNV. For continued transmission of the virus, sufficient vectors must feed on the infectious host to ensure that some mosquitoes survive long enough to feed again on a new reservoir host.

While infection in horses, dogs, and cats has been documented, these animals are not known to develop infectious-level viremias, and are therefore likely to be dead-end hosts. WNV does not appear to cause severe illness in either dogs or cats. Cases of WNV disease in horses have been documented using detection of WNV-neutralizing antibodies. In contrast to dog and cat infection, approximately 40% of equine WNV cases result in the death of the horse.

Pathogenesis and symptoms. WNV is a member of the family of Flaviviruses, a group of enveloped viruses that contain ribonucleic acid (RNA) as their nucleic acid and include Yellow fever virus, St. Louis encephalitis virus, and Dengue fever viruses. Flaviviruses are typically associated with mild systemic disease, but can cause severe infections including encephalitis. Mosquitoes acquire the flaviviruses through feeding on the blood of a viremic vertebrate host. The virus then travels from the midgut of the mosquito to the salivary glands, where it reproduces itself in great numbers. The salivary glands release virus into the saliva, where it is transferred to the next victim during a blood meal. On biting a host, the mosquito regurgitates the virus-containing saliva into the victim's bloodstream. Upon entering the host, WNV travels through the host's plasma to tissue sites for which it has a "tropism," or preference. The initial symptoms include fever, chills, headache, and backaches caused by the resulting viremia. Most viral infections do not progress beyond this stage; however, a secondary viremia may occur, which can result in infection of organs such as the brain, liver, and the skin. This can lead to fatality in the infected host.

The febrile illness associated with WNV infection is frequently nondescript, although a rash and arthritis may manifest in some patients 3 to 14 days after exposure. Lymphadenopathy (enlargement or disease of lymph nodes) may be prominent in West Nile fever. During the acute stage, the brain membranes may become involved, although this rarely leads to disease in most individuals. Typically, fatal encephalitis occurs in the elderly and in younger people with underlying illnesses, such as diabetes, emphysema, acquired immune deficiency syndrome (AIDS), and chronic diseases of the liver, kidneys, or heart. Only one in five people infected with WNV show symptoms, and less than 1 in 100 develop serious complications from the infection.

Diagnosis. WNV infection in a patient is presumptively diagnosed based on clinical symptoms and the patient's medical history. Laboratory testing is required for a confirmed diagnosis, involving a test for antibodies that are produced very early in the infected person. These antibodies can be measured in blood or cerebrospinal fluid. This blood test is generally positive in most infected people within 8 days of onset of symptoms. The current tests for human serologic diagnosis of WNV are the Centers for Disease Control and Prevention (CDC)-defined IgM and IgG enzyme-linked immunosorbent assay (ELISA). An ELISA uses antigens or "pieces" of the WNV to detect the presence of immunoglobulins or antibodies that may be present in a patient's serum (the liquid component of blood). When these antibodies are present, a positive ELISA test will result. However, because the ELISA can cross-react with other flaviviruses (such as St. Louis encephalitis virus), it is considered a screening test only and a virus

neutralization test must be conducted to confirm that WNV is indeed the virus present in the patient.

Control and prevention. Prevention and control of arbovirus disease is best accomplished through a comprehensive mosquito management program; personal, household, and community prevention through public education; and targeted prevention programs aimed at educating high-risk groups.

Mosquito surveillance and reduction. Mosquito management programs must be established at the local level and be capable of surveillance sensitive enough to detect WNV transmission associated with increased risk of disease in humans or domestic animals. Effective mosquito control includes a sustained surveillance program to detect vector species and to identify and map their immature habitats by season. Larval mosquito surveillance is performed by sampling a wide range of aquatic habitats for the presence of pest and vector species. Adult mosquito surveillance is used to monitor species presence and relative abundance of adult mosquitoes in an area.

The detection of the *Culex* mosquito in an area indicates the possible presence of WNV as well. Traps for mosquitoes use carbon dioxide as bait (because mosquitoes are attracted to the carbon dioxide that humans exhale; thus, the traps simulate a living human). After capture, WNV can then be detected in individual mosquitoes. Source reduction of mosquitoes is the elimination of mosquito larval habitats. This is the most effective method for controlling mosquitoes, and can be as simple as proper disposal of used tires, cleaning of rain gutters, bird baths, and unused swimming pools. Sanitation is a major part of integrated vector management programs, and is central to controlling WNV infection in human and mammal populations.

Personal, community, and household prevention. Individuals can use DEET (*N*,*N*-diethyl-*m*-toluamide or *N*,*N*-diethyl-3-methylbenamide)-based repellents on skin and clothing as their first defense against WNV infection, in addition to avoiding the outdoors during prime mosquito-biting hours (dusk to dawn). Households may be protected by the elimination of mosquito-breeding sites and the repair or installation of window screens. Communities can protect against WNV disease through active reporting of dead birds and organized mosquito control measures.

Targeted prevention. Targeted prevention includes seeking out those groups that are most at-risk for severe disease with WNV, including those over age 50 and those who are immunocompromised. Individuals who work outside extensively or engage in recreation outdoors in endemic areas are also at greater risk for WNV infection and should be also targeted with disease-preventing information.

Treatment and preventive vaccine. Treatment for WNV encephalitis and fever is largely supportive: possible hospitalization, administration of intravenous fluids, respiratory support (ventilation), and prevention of secondary infections. There is no specific antiviral therapy for WNV infection.

No vaccine is available for use to prevent WNV disease in humans. One research group has successfully used recombinant deoxyribonucleic acid (DNA) vaccines to induce protective immune responses in mice and horses, but this success has yet to be transferred to humans. *See* ARBOVIRAL ENCEPHALIDES; MOSQUITO; VIRUS; ZOONOSES.

Marcia M. Pierce

Bibliography. R. W. Bauman, *Microbiology: Alternate Edition with Diseases by Body System*, Benjamin Cummings, 2006; P. R. Murray et al., *Medical Microbiology*, Mosby, 2002; E. W. Nester et al., *Microbiology: A Human Perspective*, 4th ed., McGraw-Hill, 2004; L. M. Prescott, J. P. Harley, and D. A. Klein, *Microbiology*, 6th ed., McGraw-Hill, 2005.

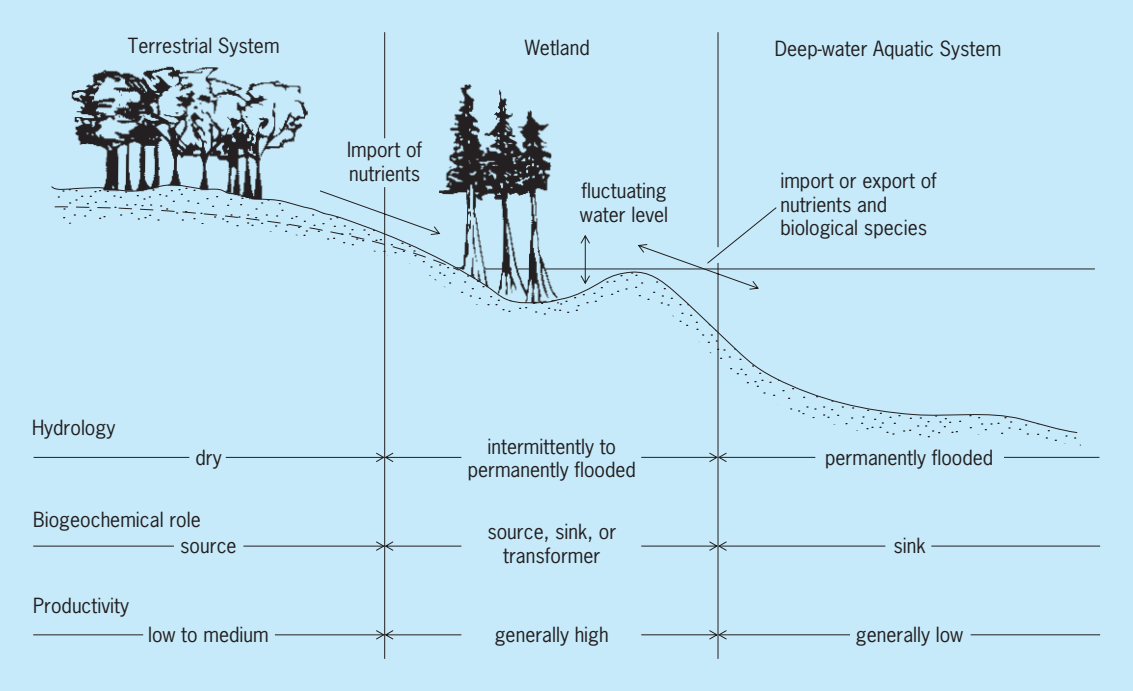
Wetlands

Ecosystems that form transitional areas between terrestrial and aquatic components of a landscape. Typically they are shallow-water to intermittently flooded ecosystems, which results in their unique combination of hydrology, soils, and vegetation. Examples of wetlands include swamps, fresh- and salt-water marshes, bogs, fens, playas, vernal pools and ponds, floodplains, organic and mineral soil flats, and tundra. As transitional elements in the landscape, wetlands often develop at the interface between drier uplands such as forests and farmlands, and deep-water aquatic systems such as lakes, rivers, estuaries, and oceans (see **illus.**). Thus, wetland ecosystems are characterized by the presence of water that flows over, ponds on the surface of, or saturates the soil for at least some portion of the year.

Features. Regulations supporting the U.S. Federal Clean Water Act [33 CFR 328.3 (b)] define wetlands as "those areas that are inundated or saturated by surface or ground water at a frequency and duration sufficient to support, and that under normal circumstances do support, a prevalence of vegetation typically adapted for life in saturated soil conditions." A more inclusive definition of a wetland was adopted by the international community through the 1971 Ramsar Convention, which states that wetlands are "areas of marsh, fen, peat, and or water, whether natural or artificial, permanent or temporary, with water that is static or flowing, fresh, brackish or salt, including areas of marine water the depth of which at low tide does not exceed six meters."

Wetland soils can be either mineral (composed of varying percentages of sand, silt, or clay) or organic (containing 12–20% organic matter). Through their texture, structure, and landscape position, soils control the rate of water movement into and through the soil profile (the vertical succession of soil layers). Retention of water and organic carbon in the soil environment controls oxidation-reduction and other biogeochemical reactions that facilitate the functioning of wetland soils. Wetland soils are regularly wet, deprived of oxygen, or both. *See* BIOGEOCHEMISTRY; OXIDATION-REDUCTION.

Vegetated wetlands are dominated by plant species, called hydrophytes, that are adapted to live in water or under saturated soil conditions.



Relationships of wetlands to terrestrial systems and deep-water aquatic systems and comparison of their characteristics. (*After W. J. Mitsch and J. G. Gosselink, Wetlands, 2d ed., Van Nostrand Reinhold, 1993*)

Adaptations that allow plants to survive in a waterlogged environment include morphological features, such as pneumatophores (the "knees," or exposed roots, of the bald cypress), buttressed tree trunks, shallow root systems, floating leaves, hypertrophied lenticels, inflated plant parts (such as aerenchyma), and adventitious roots, to name a few. Physiological adaptations also allow plants to survive in a wetland environment. These adaptations include the ability of plants to transfer oxygen from the root system into the soil immediately surrounding the root (rhizosphere oxidation); the reduction or elimination of ethanol accumulation due to low concentrations of alcohol dehydrogenase (ADH); and the ability to concentrate malate (a nontoxic metabolite) instead of ethanol in the root system. *See* ROOT (BOTANY).

Wetlands differ with respect to their origin, position in the landscape, and hydrologic and biotic characteristics. For example, work in the United States has focused on the hydrology (water sources and transport processes) as well as the geomorphic position of wetlands in the landscape. This hydrogeomorphic approach, developed by M. M. Brinson and colleagues, recognizes and uses the fundamental physical properties that define wetland ecosystems to distinguish among classes of wetlands that occur in riverine, depressional, estuarine or lake fringe, mineral or organic soil flats, and slope environments. Ecosystem functions (such as subsurface water storage) in the broad categories of hydrology, biogeochemistry, plant community, and faunal support/habitat are specific to each class of wetland (see **Table 1**).

Global extent. The extent of wetlands in the world is estimated to be 2–3 $\times$ 10^6 mi^2 (5–8 $\times$ 10^6 km^2), or about 4–6% of the Earth's land surface. Wetlands occur in humid, cool regions such as bogs, fens, and tundra of Canada, Alaska, and northern Europe and Asia; along rivers and streams as riparian wetlands, seasonally flooded forests, and backswamps; in the deltas of the great rivers, including the Amazon, Mekong, Danube, Mississippi, Nile, and Rhone; along temperate, subtropical, and tropical coastlines as salt marshes, mud flats, and mangrove swamps; and even in arid regions as inland salt flats,

TABLE 1. Wetland functions typical of riverine ecosystems in North America as identified by the hydrogeomorphic approach*

Functional group	Function
Hydrology	Dynamic surface water storage
	Long-term surface water storage
	Energy dissipation
	Subsurface storage of water
	Ground-water flow or discharge
Biogeochemistry	Nutrient cycling
	Removal of elements and compounds
	Retention of particulates
	Organic carbon export
Plant community	Maintenance of native plant community
	Detrital biomass production
Faunal support/habitat	Spatial structure of habitat
	Interspersion and connectivity
	Distribution and abundance of native invertebrate fauna
	Maintenance of the distribution and abundance of native verbrate fauna

*From M. M. Brinson et al., *Guidebook for Application of Hydrogeomorphic Assessments to Riverine Wetlands*, Tech. Rep. WRP-DE-11, U.S. Army Corps of Engineers Waterways Experiment Station, Vicksburg, MS, 1995.

seasonally inundated playas, riparian systems, and vernal pools. Wetlands are found on every continent except Antarctica and in every clime from the tropics to the frozen tundra. Rice paddies, which comprise another 500,000–600,000 mi^2 (1.3–1.5 × 10^6 km^2), can be considered as a type of domesticated wetland of great value to human societies worldwide. In the contiguous United States, an estimated 160,000 mi^2 (420,000 km^2) of wetlands remain, with 490,000 mi^2 (1,270,000 km^2) in Canada and 266,000 mi^2 (688,000 km^2) in Alaska. Most Canadian and Alaskan wetlands are considered northern peatlands (containing histosols, or soils dominated by organic material). *See* BOG; MANGROVE; MUSKEG; PLAYA; SALT MARSH; TUNDRA.

Societal values. Compared to other ecosystems, wetlands are often an extremely productive part of the landscape. Wetlands support a rich variety of waterfowl and aquatic organisms, and represent one of the highest levels of species diversity and richness of any ecosystem. From a conservation viewpoint, wetlands are an extremely important habitat for rare and endangered species. For example, over 63% of reptiles, 68% of birds, and 75% of amphibians listed as threatened or endangered in the United States are associated at some point in their life cycle with wetlands.

Wetlands often serve as natural filters for human and naturally generated nutrients, organic materials, and contaminants. The ability to retain, process, or transform nutrients, organic matter, and contaminants is called assimilative capacity, and is strongly related to wetland soil texture and vegetation. The assimilative capacity of wetlands has led to many projects that use wetland ecosystems for wastewater treatment and for improving water quality. Wetlands have been constructed to process domestic sewage, coal mine drainage, industrial wastes, urban runoff, and agricultural drainage waters. Wetlands also have been shown to prevent downstream flooding and, in some cases, to prevent ground-water depletion as well as to protect shorelines from storm damage. *See* GROUND-WATER HYDROLOGY.

Management. The best wetland management practices enhance the natural processes of wetlands by maintaining conditions as close to the natural hydrology of the wetland as possible, including hydrologic connections with adjacent rivers, lakes, and estuaries. Marsh management for wildlife, particularly waterfowl, often has translated to water-level manipulation achieved by the construction of dikes, weirs, control gates, and pumps.

Losses. Wetlands can be easily affected by drainage projects and upstream hydrologic modifications. As a result, the world's wetlands are becoming a threatened landscape. Loss of wetlands worldwide currently is estimated at 50%. Although in a few regions of Scandinavia, North America, Africa, and Australia some natural wetlands still can be found in large tracts, in some countries, especially in Europe, natural wetlands are nearly extinct or, at best, heavily managed. It has been estimated that the contiguous United States has lost 53% of its wetlands over the past 200 years, with several states in the upper Midwest losing 80% or more. Wetlands in California have been reduced, according to some estimates, to less than 10% of their original distribution.

Wetland loss results primarily from habitat destruction, alteration of wetland hydrology, and landscape fragmentation (**Table 2**). Global warming may soon be added to this list, although the exact loss of coastal wetlands due to sea-level rise is not well documented. Worldwide, destruction of wetland ecosystems primarily has been through the conversion of wetlands to agricultural land. For example, over 80% of the Mississippi riverine bottomlands, and well over 90% of historic wetlands in Ohio, have been converted to crop production. Other wetland destruction has occurred through the harvesting of timber. The practice of clear cutting and the construction of logging roads are two major forestry activities that directly destroy wetlands, although degradation

TABLE 2. Human actions leading to wetland loss*

Cause of loss	Floodplains	Rivers	Lakes	Peatlands	Swamps
Drainage for agriculture, forestry, and mosquito control	C	C	P	C	C
Dredging and channelization for navigation and flood protection	A	P	A	A	A
Filling for solid waste disposal, roads, and commercial, industrial, or residential development	P	P	P	A	A
Conversion for aquaculture	C	C	P	A	A
Construction of dikes, dams, and seawalls for flood and storm control, water supply, and irrigation	C	C	C	A	A
Discharge of pesticides, herbicides, domestic and industrial waste. agricultural runoff, and sediment	C	C	C	A	A
Mining of wetlands for peat, coal, gravel, phosphate, and other materials	P	A	C	C	C
Logging and shifting cultivation	C	P	A	C	C
Ground water abstraction	P	C	A	A	A
Fire	C	P	A	C	C
Sediment diversion by dams, deep channels, and other structures	C	C	P	A	A
Hydrological alteration by canals, roads, and other structures	C	C	C	P	P
Subsidence due to extraction of ground water, oil, gas, and other minerals	C	C	A	A	A

*C = common and important cause of loss. P = present but not a major causes of loss. A = absent or exceptional.
SOURCE: From P. Harrison and F. Pearce, *AAAS A has of Population and Environment*, University of California Press, Berkeley, 2001; based on data of United Nation Environment Programme.

through other forms of timber extraction also is widespread. Peat extraction for fuel and horticulture occurs primarily in the Northern Hemisphere (Great Britain, North America, and northern Eurasia), and is responsible for the loss of millions of hectares of peatlands. Other worldwide activities that have led to the direct loss of wetlands include sand and gravel mining of riverbeds and banks and the development of mariculture (cultivation of marine species) through the conversion of salt marshes and mangrove forests into commercial impoundments.

Hydrologic modifications that destroy, alter, and degrade wetland systems include the construction of dams and water diversions, ground-water extraction, and the artificial manipulation of the amount, timing, and periodicity of water delivery. The primary impact of landscape fragmentation on wetland ecosystems is the disruption and degradation of wildlife migratory corridors, reducing the connectivity of wildlife habitats and rendering wetland habitats too small, too degraded, or otherwise irreversibly altered to support the critical life stages of plants and animals.

Protection policies. The heavy losses of wetlands in the world, coupled with the recognized values of these systems, have led to a number of policy initiatives, especially in the United States, where estimated losses have been as high as 300,000 acres (1200 km^2) per year. The federal government has no specific wetland protection law. Instead, it has relied primarily on presidential executive orders, a "no net loss" policy, and Sections 404 and 401 of the Clean Water Act of 1977 for wetland protection. The principal agencies involved in Clean Water Act jurisdiction include the Environmental Protection Agency, Army Corps of Engineers, Fish and Wildlife Service, and National Marine Fisheries Service. The efforts of these regulatory agencies with Clean Water Act jurisdiction have been augmented by a wetland protection program in agriculture ("Swampbuster" Provisions of the Food Security Act) administered by the Natural Resources Conservation Service of the Department of Agriculture. Swampbuster provisions deny federal subsidies to any farm owner who converts certain types of wetlands to farmland. Canada, Sweden, and other countries also have developed wetland protection regulations.

International cooperation. International agreements, particularly the Ramsar Convention and the North American Waterfowl Management Plan, govern the protection of wetland ecosystems worldwide. The Ramsar Convention is an international treaty for the conservation of wetland ecosystems. Named for the city in Iran where the Convention on Wetlands took place in 1971, the Ramsar Convention maintains a list of "Wetlands of International Importance." Inclusion on the list is determined by the international significance of a wetland, with an emphasis on the its ability to support migratory waterfowl. Africa and North America hold the largest percentage of Ramsar wetlands (21% and 23%, respectively), while South America supports another 16%. As of 2000, a total of 66,840,000 hectares in more than 890 sites is protected under the Ramsar Convention.

The North American Waterfowl Management Plan, agreed to in 1986 by the United States and Canada, has a goal of conserving and restoring about 9300 mi^2 (24,000 km^2) of waterfowl wetland habitat. This treaty was developed partly in response to a steep decline in waterfowl in both countries in the early 1980s.

Restoration and creation. Wetland restoration usually refers to the rehabilitation of degraded or hydrologically altered wetlands, often involving the reestablishment of vegetation. Wetland enhancement generally refers to the targeted restoration of one or a set of ecosystem functions over others, for example, the focused restoration of a breeding habitat for rare, threatened, or endangered amphibians. Wetland creation refers to the construction of wetlands where they did not exist before. This type of activity typically involves much more extensive hydrology and soils engineering. Created wetlands are also called constructed or artificial wetlands. Some regulatory agencies distinguish between constructed wetlands and created wetlands, with the former term used to designate wetlands built for wastewater or storm-water treatment, and the latter term used for wetlands developed on nonwetland sites to produce or replace a natural habitat.

Restoring, enhancing, or creating a wetland requires a comprehensive understanding of hydrology and ecology, as well as engineering skills. Failure of wetland creation and restoration usually has resulted from one or more of the following circumstances: poor understanding of the restored site's hydrology: lack of appreciation for the intrinsic dynamics of wetlands and their ecology; unrealistic expectations as to what constitutes a successfully created or restored wetland ecosystem; and not allowing enough time for the wetland ecosystem to develop.

There is a great deal of interest in wetland restoration and creation, particularly in response to the "no net loss" policy in the United States that requires replacement of wetlands when they are unavoidably filled for development. In addition, thousands of wetlands have been constructed throughout the world as part of an economical and ecological approach to improving water quality, especially for domestic wastewater, non-point-source pollution from agriculture, and coal mine drainage. Although it is difficult to estimate the extent of created and restored wetlands in the world, rates of wetland gain probably do not equal the rates of wetland loss. *See* DAM; ECOSYSTEM; ESTUARINE OCEANOGRAPHY; HYDROLOGY; RESTORATION ECOLOGY; RIVER ENGINEERING.

William J. Mitsch; Peggy L. Fiedler; Lyndon C. Lee; Scott R. Stewart

Bibliography. B. L. Bedford, D. J. Leopold, and J. P. Gibbs, Wetland ecosystems, in S. A. Levin (editor in chief), *Encyclopedia of Biodiversity*, vol. 5, Academic Press, New York, 2001; C. M. Finlayson and M. Moser (eds.), *Wetlands*, 1991; D. A. Hammer, *Creating Freshwater Wetlands*, 2d ed., Lewis Publishers, 1996; D. A. Hammer (ed.), *Constructed Wetlands*

for Wastewater Treatment, Lewis Publishers, Chelsea, MI, 1989; P. Harrison and F. Pearce, *AAAS Atlas of Population and Environment*, University of California Press, Berkeley, 2001; R. H. Kadlec and R. L. Knight, *Treatment Wetlands*, Lewis Publishers/CRC Press, Boca Raton, FL, 1995; E. Maltby, *Waterlogged Wealth*, Earthscan Publishers, London, 1986; W. J. Mitsch (ed.), *Global Wetlands: Old World and New*, Elsevier, Amsterdam, 1994; W. J. Mitsch and J. G. Gosselink, *Wetlands*, 3d ed., Wiley, New York, 2000; T. C. Winter, Hydrologic studies of wetlands in the northern prairie, in A. van del Valk (ed.), *Northern Prairie Wetlands*, Iowa State University Press, Ames, 1989.

Wheat

A food grain crop. Wheat is the most widely grown food crop in the world, and is increasing in production. It ranks first in world crop production and is the national food staple of 43 countries. At least one-third of the world's population depends on wheat as its main staple. The principal food use of wheat is as bread, either leavened or unleavened.

The United States is second to Russia in total production, but the average yield per acre in the United States is about twice that of Russia. Other major wheat-producing countries in the world are Canada, China, India, France, Argentina, and Australia. Highest per acre yields (60 bushels or 1.63 metric tons, and above) are produced in western Europe. The change to day-length-insensitive, short-straw and stiff-straw varieties which can be heavily fertilized and irrigated has resulted in yield increases, especially in food-deficit countries from about 35°N latitude to the Equator (**Fig. 1**). Wheat is an important commodity of international trade. Of the total wheat production, about one-fifth moves in world trade. The United States leads in exports. Other leading export countries include Canada, Australia, France, and Argentina.

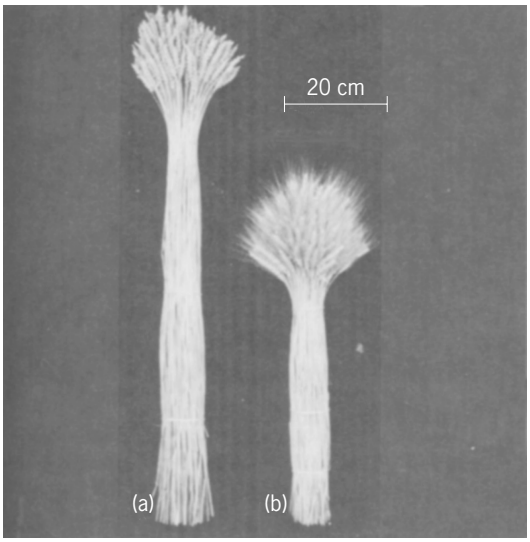


Fig. 1. Sheaves of wheat: (*a*) tall, awnless variety; (*b*) short, awned or bearded variety. Both tall and short types may have awns.

Adaptation. Wheat is best adapted to a cool dry climate, but is grown in a wide range of soils and climates. Much of the world's wheat is seeded in the fall season and, after being dormant or growing very slowly during winter, it makes rapid growth in the spring and develops grain for harvest in early summer. In the United States, about 80% of the wheat is fall-seeded. Wheat competes with other crops where the fall-seeded type makes good use of limited rainfall and where the spring-seeded type makes good use of the short frost-free growing season. Medium-textured and well-drained soils are best for wheat since it cannot withstand long periods of wet ground.

General cultural practices. Clean, weed-free seed of an adapted variety treated to control seed-borne diseases, along with good germination, increases the prospect of a satisfactory crop. Wheat is usually seeded at a depth of 1–2 in. (2.5–5 cm), or deeper when topsoil is dry. A firm seed bed, together with soil firmly packed over the seed, improves stands and reduces losses from winter kill and similar hazards. Wheat may be seeded as part of a rotation following corn, soybeans, or other small grain crops in humid regions; a year of fallow during which moisture is accumulated and weeds kept under control often precedes wheat in areas of low rainfall. Seeding time depends on soil moisture, soil temperature, presence of certain insects and disease organisms, and other factors. Seeding rate averages about 60 lb/acre (67 kg/hectare) in the United States but may be as high as 120 lb/acre (134 kg/ha) in the eastern United States and as low as 30 lb/acre (33 kg/ha) in the arid West.

Most wheat is harvested directly from the field with a combined harvester-thresher, or it may be wind-rowed first and then combined when dry. Wheat grain should contain no more than 13% moisture for safe storage, should be cool, and free from insects. Forced-air dryers are used to reduce the moisture content of grain harvested at above 13%. *See* AGRICULTURAL MACHINERY.

Market classes. Wheat for milling is classified according to hardness, color, and best use. In the United States, there are seven official market classes of which the following five are the most important: (1) hard red winter, for bread; (2) hard red spring, for bread and rolls; (3) soft red winter, for cake and pastries; (4) white, for bread, breakfast foods, and pastries; and (5) durum, for macaroni products.

Distribution of types and varieties. Variety adaptation, production practices, and crop hazards are widely different in the major areas of production in the United States. Therefore, the United States can be divided into six wheat-growing regions (**Fig. 2**), although the boundaries overlap: region I, the Northeast and Ohio Valley; region II, the Southeast; region III, the Northern Plains; region IV, the Central Plains; region V, the Pacific Northwest; and region VI, California and Arizona.

Regions I and II have the highest rainfall of the six, and wheat of soft texture with low protein

generally is grown. The crop is fall-seeded. General crop rotations are used in which corn and soybeans take prominent places. Yorkstar, Genesee, and Ionia, all white grains, are the leading varieties in region I. In region II, Arthur, a soft red winter type, is the leading variety in most of the states.

In region III, spring-sown wheat predominates. The leading varieties of durum wheat are Rolette and Leeds, while the leading hard red spring varieties are Era, Waldron, and Olaf. Some fall-sown wheat is grown in the western portion of this region, but the winters are severe and the crop often is lost. In general, the region has low rainfall; hence tillage and rotations are designed to conserve as much moisture as possible.

In region IV, little else but hard red winter wheat is grown. Subregion IVA requires a cold-hardy variety, and subregion IVB, with its hot summers, requires an early maturing variety for most dependability. In subregion IVA, leading varieties in recent years have been Winalta and Cheyenne; in subregion IVB, Scout, Centurk, Triumph, and Sturdy. This region has rather low rainfall, and tillage, including summer fallow, is especially timed to conserve moisture for the next crop.

In region V, wheat of several types is grown but soft white and hard red winter predominate. Much of region V uses summer fallow in the cropping sequence. A small acreage is irrigated. Nugaines and Hyslop have been the most important white wheat varieties in this region. Paha and Moro, club wheats with white grain, have been grown also. Wanser is the leading hard red winter variety, and Komar and Thatcher hard red spring wheats continue to be found on a small acreage.

Region VI grows day-length-insensitive common wheats with red or white grain, or amber durum. Irrigation is common for wheat in this region. The mild winters allow nonhardy varieties to be utilized from fall seeding. The leading varieties in recent years have been Anza and Inia 66.

In all of these regions, variety changes occur every year as improved forms are introduced; hence, the varieties named here represent types but may not continue to be the particular varieties actually produced in a given year. Some wheat is irrigated in regions IV, V, and VI, but the major acreage is rainfed.

Northwestern Mexico grows varieties similar to those found adapted in region VI and utilizes similar production methods. In Canada the varieties used are similar to those used in region III. Production practices in Canada are also similar to those used in region III, with heavy emphasis on dry-land farming. The leading variety of hard red spring wheat in Canada is Neepawa.

Characteristics. The wheat inflorescence is a spike bearing sessile spikelets arranged alternately on a zigzag rachis. Two, three, or more florets may develop in each spikelet and bear grains. The florets are composed of three stamens and a single ovary with its style and stigma enclosed by a lemma and palea (**Fig. 3**). The lemma may be awned or awnless.

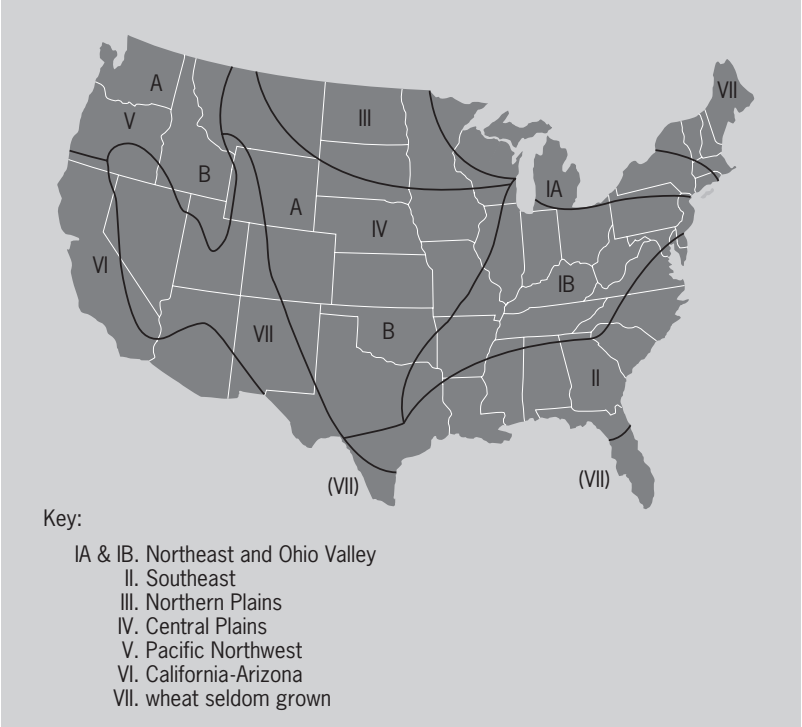


Fig. 2. Adaptation regions for wheat in the United States. (*USDA*)

The grain may be white, red (brown), or purple, and it may be hard or soft in texture. Size of the grain or caryopsis may be large, as in durum, or very small, as in shot wheat (*Triticum sphaerococcum*). Wheats vary in plant height and in the ability to produce tillers. The stems are usually hollow.

Wheat is normally a self-pollinating plant, although out-crosses do occur. When a breeder deliberately makes crosses, the anthers must be removed before

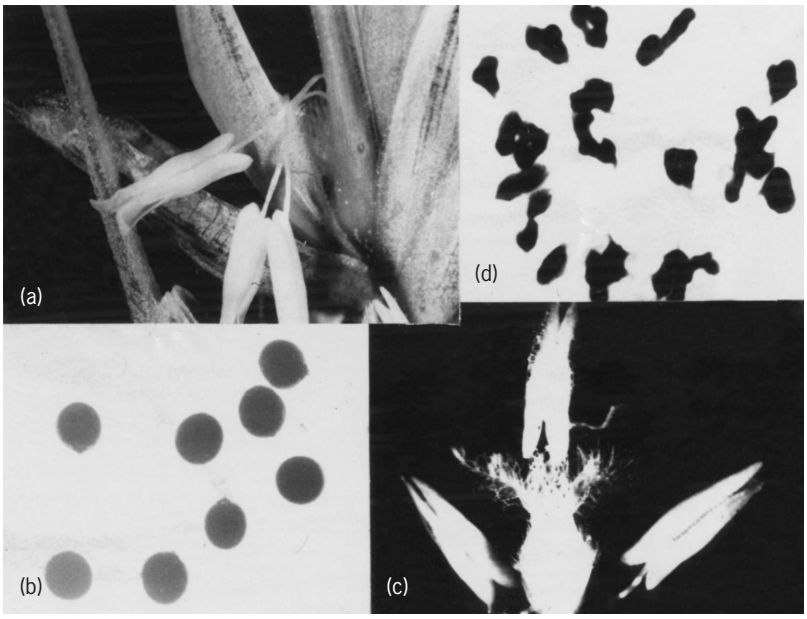


Fig. 3. Flower parts of a male fertile wheat plant: (*a*) three stamens; (*b*) pollen grains; (*c*) single ovary with feathery stigma; and (*d*) normal chromosomes.

they shed pollen, and fresh pollen must be transferred from the second parent in the cross to the stigmas while they are receptive. Certain breeding stocks of wheat have been found that are male-sterile and shed no pollen. These have been used successfully to develop hybrid wheats without resorting to hand pollination procedures.

The wheat grain is composed of the endosperm and embryo enclosed by bran layers. The endosperm portion is principally starch and is therefore used as energy food. Wheat is also an important protein source, especially for those people who use wheat as their main staple. Since wheat protein is deficient in lysine and some of the other essential amino acids that are important in human nutrition, breeding for improved amino acid balance or nutritional value is a primary objective in many breeding projects. *See* SEED.

Origin and relationships. Wheat was gathered for food in prehistoric times and has been grown as a crop for perhaps 7000 years. The exact center of origin of cultivated wheat has not been established, but the most likely area is the foothills of southwestern Iran, southeastern Turkey, and adjacent parts of Iraq and Syria with an extension to Israel. Wild *Aegilops* and *Triticum* are still found in these regions.

Botanically, wheat is a member of the grass family to which rice, barley, corn, and several other cereal grain crops also belong. The *Triticum* genus includes a wide range of wheat forms (**Fig. 4**). They are often grouped by their polyploid chromosome number series of 7, 14, and 21 pairs. In Fig. 4, *a* and *b* have 7 pairs (diploid), *c* to *l* have 14 pairs (tetraploid), and *m* to *q* have 21 pairs (hexaploid) of chromosomes.

Taxonomic studies place the goat grasses (*Aegilops*) and wheat (*Triticum*) in one genus, *Triticum*. They intercross to some extent, and it is believed that *Aegilops* species have contributed one or more groups of chromosomes to the tetraploid and hexaploid species of *Triticum*. Breeders have transferred individual genetic characteristics from goat grasses to wheat. The forms illustrated in Fig. 4 have been intercrossed, and in this way resistance to wheat rusts has been transferred from durum and emmer to common wheat. Wheat has been crossed with rye (*Secale*) and with *Agropyron* (a grass), and by the use of irradiation, segments of alien chromatin have been transferred to wheat, sometimes with beneficial results. New forms, called *Triticale*, have been derived from crossing rye and wheat followed by doubling the chromosomes in the hybrid. The most successful of these is based on a cross of rye and durum wheat in which there are 14 pairs of wheat and 7 pairs of rye chromosomes in the new species. Several variations occur.

Since there is much genetic similarity among the three genomes (sets of 7) of common wheat, the loss of a chromosome of one genome is partially compensated by similar chromosomes in the other genomes. This allows the development of 21 different genetic stocks, each deficient for one chromosome or one chromosome pair. The use of such stocks has greatly accelerated genetic studies and has made work more precise. An extension of this technique permits the substitution of a whole chromosome in one variety for the same chromosome in another variety. This has value in genetic studies and is also useful in variety improvement. *See* BREEDING (PLANT).

Fig. 4. Spikes of some varieties of wheat. (*a*) Wild and (*b*) cultivated forms of einkorn (*Triticum monococcum*); (*c*) wild emmer (*T. dicoccoides*); (*d*) emmer (*T. dicoccum*); (*e*) durum (*T. durum*); (*f*) poulard (*T. turgidum*); (*g*, *h*) durum (*T. durum*); (*i*) Polish (*T. polonicum*); (*j*) Persian (*T. carthlicum*); (*k*) emmer (*T. dicoccum*); (*l*) timopheevi (*T. timopheevi*); (*m*) spelt (*T. spelta*); (*n*) bread or common (*T. aestivum*); (*o*) club (*T. compactum*); (*p*) shot (*T. sphaerococcum*); and (*q*) spelt (*T. spelta*). (*Courtesy of H. Kihara*)

Variety development. Most countries in which wheat is grown have wheat breeding programs in which the objective is to develop more productive and more stable varieties (cultivars). Many methods are combined in these programs, but in nearly all of them specially selected parent types are crossbred followed by pure-line selection among the progeny to develop new combinations of merit. Varieties and genetic types from all over the world become candidate parents to provide the desired recombinations of good quality, winter and drought hardiness, straw strength, yield, and disease resistance. Wheats must be bred for specific milling processes and to provide quality end-use products. Many new varieties have complex pedigrees.

A number of worldwide collections of diverse wheats are available to breeders. The collection maintained by the U.S. Department of Agriculture contains about 35,000 items. Teams of breeders, geneticists, entomologists, plant pathologists, and

cereal chemists are found at major breeding centers to give comprehensiveness to the breeding of the new varieties. *See* GRAIN CROPS. Louis P. Reitz

Diseases

Estimates of annual loss of wheat due to disease range from 10 to 20%. More than 100 diseases of wheat have been described, and about 50 may be economically important. Stem rust, leaf rust, loose smut, and stinking smut are potentially destructive and have caused enormous losses in the past. Currently these diseases are better controlled. Virus diseases and the root and basal stem rots are widespread and cause significant loss.

The loss in North America from black stem rust, caused by the fungus *Puccinia graminis* f. sp. *tritici*, in earlier years has been as high as 3×10^8 bu (1.1×10^7 m^3). This fungus is heteroecious, with its sexual phase and aeciospores produced on the common barberry (*Berberis vulgaris*). The aeciospores from barberry germinate and invade wheat stems and leaves, producing urediniospores in reddish-brown pustules (**Fig. 5**) and later teliospores in black pustules. The teliospores over-winter and then germinate in the spring, producing basidiospores which initiate the life cycle on the barberry in areas where this shrub is present. Stem rust can also survive over winter as urediniospores on wheat in southern Texas, Mexico, and other countries with mild climates. Spread by airborne urediniospores into more northern wheat growing areas then occurs during the spring and summer. Eradication of the common barberry in much of North America has helped to reduce losses from stem rust by eliminating the sexual stage, preventing the development of new pathogenic races by hybridization, and also by eliminating aeciospores, one of the sources of inoculum for wheat.

Fig. 5. Black stem rust of wheat urediniospore stage.

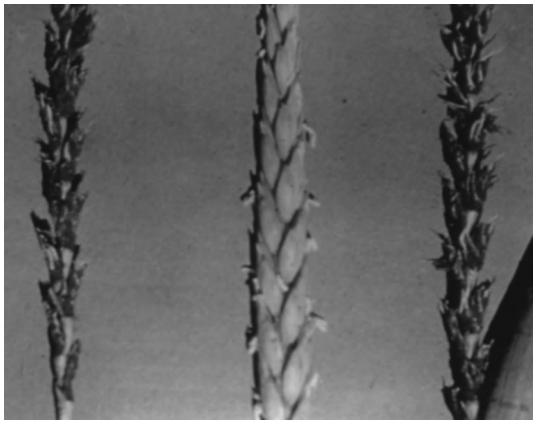


Fig. 6. Loose smut of wheat-diseased and healthy spikes.

Losses from stem rust and from the leaf rust fungus, *Puccinia recondita*, have been reduced in recent years by using cultivars of wheat with genetic resistance to the prevalent physiologic or pathologic races of both pathogens.

The common smut diseases of wheat are loose smut, caused by *Ustilago nuda*, and stinking smut, caused by *Tilletia* species. Both fungi are systemic in their host. At flowering the entire head (spike) is transformed into a mass of smut spores in the loose smut disease (**Fig. 6**). Spores develop where the grains usually form in stinking smut, but the accessory parts of the spike are not affected. The diseased heads therefore are not as conspicuous as in loose smut. An odor resembling decaying fish is associated with stinking smut. Both fungi are seed-borne, as spores externally in case of stinking smut and as mycelium within the healthy grain in the loose smut disease. Spores of stinking smut also survive in the soil in some areas. When seed with smut spores or with internal mycelium are sown, the smut fungi grow into the young seedlings, but usually there are no distinct symptoms until heading. Chemical seed treatment fungicides are effective controls of the smut diseases except where the spores survive in the soil. Resistant cultivars are also available for some areas of wheat culture. *See* FUNGISTAT AND FUNGICIDE.

Since 1950, several viral diseases have become common and destructive. The symptoms of these diseases include mosaic and mottling of foliage, stunting, rosetting, and yellowing. The virus causing wheat streak mosaic is transmitted by the mite *Aceria tulipae*. Other grains and grasses are also susceptible. The vectors of barley yellow dwarf virus are several species of aphids. Wheat spindle streak and soil-borne mosaic are caused by soil-borne viruses that are transmitted by *Polymyxa graminis*, a soil-borne fungus. Zoospores of *Polymyxa* with viral particles invade wheat roots, thereby effectively inoculating the plants. Late planting, crop rotation, and the use of resistant cultivars are some of the management practices that aid in reducing losses from these

diseases.

There are several damaging root, crown, and basal stem rot diseases in many wheat-growing areas of the world. Stunting, early kill of plants, lodging, and premature ripening are the symptoms. The causal fungi include *Fusarium* species, *Gaeumannomyces graminis*, *Pseudocercosporella herpotrichoides*, and *Bipolaris sorokiniana*. These fungi persist in the soil or crop residues from year to year. Crop rotation, late planting of winter wheat, and ensuring adequate soil fertility are management practices that reduce losses. *See* PLANT PATHOLOGY.

C. Wayne Ellett

Wheat Grain Proteins

Most of the protein of wheat grain is found in the endosperm, the inner part of the kernel that yields white flour in the milling process. The endosperm—about 75% of the wheat kernel—consists largely of starch granules surrounded by a proteinaceous matrix. The protein makes up about 8–16% of the grain, but lower or higher protein percentages are not unusual. *See* PROTEIN.

Most of the protein in the endosperm is associated with gluten, the cohesive and elastic protein mass that remains after a dough is kneaded in water or in an aqueous salt solution to wash away the starch granules. The ability to form gluten is unique to wheat (although the term "corn gluten" has been applied somewhat inappropriately in recent years to a maize protein fraction that lacks gluten-like viscoelastic properties). Gluten proteins are storage proteins, which are broken down to provide nitrogen and amino acids for the synthesis of proteins and other molecules needed by the new plant upon seed germination. In addition to gluten, endosperm proteins include 15–25% of proteins that have been classified as albumins or globulins based on their solubilities. Many of these proteins may, like gluten, serve as storage proteins, although various proteins of the albumin-globulin fraction inhibit protease, amylase, or other enzymes. These inhibitor proteins may protect the grain against insect or fungal attack. Various other proteins occur in the bran (outer layer) or germ (embryo) fractions obtained during milling, but these constitute a small part of the total protein. *See* ALBUMIN; GLOBULIN.

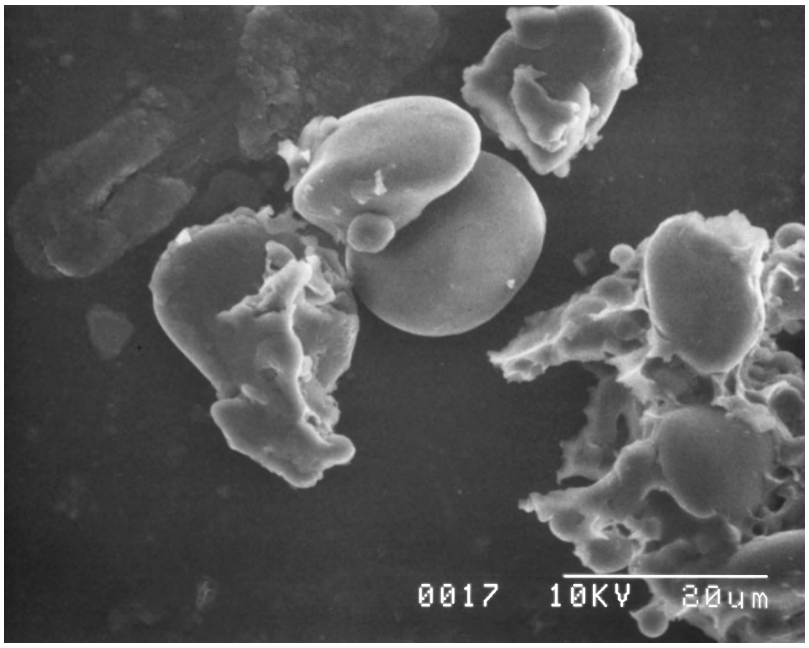


Fig. 7. Scanning electron micrograph of flour particles from a hard red winter wheat cultivar (Len). Large and small spherical objects are starch granules. The irregularly shaped material attached to the starch granules corresponds to protein deposits that surround the starch granules in the endosperm cells. (*Micrograph by Gregory M. Glenn, USDA, Albany, CA*)

During grain development when storage protein and starch synthesis occur, storage proteins are deposited in the endosperm as protein bodies. These protein bodies and any membranes surrounding them are disrupted during the desiccation phase that leads to the dry mature grain. Consequently, the protein ends up as a largely continuous matrix surrounding the starch granules. Adherence of the endosperm to the surrounding protein matrix differs in hard and soft wheats. In soft-milling cultivars, the matrix separates more readily from the granules, generally leaving them intact; however, in hard-milling cultivars, the matrix and granules adhere so strongly that starch granules are occasionally cleaved during milling. For large-scale flour milling, machinery must be adjusted appropriately for grain hardness in order to achieve good yields of white flour (endosperm). Soft wheats are commonly used for pastry or noodle making, whereas the hard wheats are more commonly used for breadmaking. **Figure 7** is a scanning electron micrograph of some flour particles from a hard red winter wheat cultivar. *See* STARCH.

Gliadin and glutenin. Traditionally, gluten proteins were divided into two main fractions on the basis of solubility (gliadin) or insolubility (glutenin) in aqueous alcohol solutions (such as 70% ethanol-water or 50% *N*-propanol–water). This is a somewhat crude fractionation, but it is now recognized that the gliadin fraction consists almost entirely of monomeric (having single polypeptide chains) proteins, and the glutenin fraction consists almost entirely of polymers, in which certain gluten proteins are linked together by disulfide bonds. The proteins of glutenin polymers are often called subunits to distinguish them from the monomeric gliadins. Thus, just as polypeptides or proteins are polymers of amino acids, glutenin is a higher-order polymer in which the monomers are proteins.

All gluten proteins have substantial amounts of proline and glutamine, 2 of the usual 20 amino acids found in proteins. Consequently, the proteins are often classified as prolamins (although in earlier work, this term was usually reserved for the gliadin proteins). Prolamins are also found in related cereal grains, including rye, barley, and corn.

Gluten can be separated by two-dimensional electrophoretic or chromatographic methods into about 40–50 components, but the actual number of components is bound to be significantly higher based on the number of gluten protein genes found in the wheat genome. Gluten proteins are about equally divided between gliadin and glutenin; however,

the actual number of components varies among cultivars. Structurally, the gliadins can be divided into α-, γ-, and ω-gliadins on the basis of their amino acid sequences. Gutenin subunits are usually divided into a high-molecular-weight group (mol wt ~70,000–100,000) and a low-molecular-weight group (mol wt ~30,000–42,000). Of all the endosperm proteins, the amounts and types of high-molecular-weight glutenin subunits show the strongest correlations with variations in mixing and baking quality among cultivars. *See* CHROMATOGRAPHY; ELECTROPHORESIS.

Breadmaking. All gluten proteins are strongly cohesive due not only to hydrogen bonding between the many glutamine residues but also to hydrophobic and ionic interactions. It is, however, the polymeric glutenin fraction that contributes elasticity to the hydrated gluten proteins, as in a wheat flour-and-water dough. Wheat cultivars vary widely in quality for breadmaking and other food uses. For breadmaking, dough should have a moderately high level of elasticity and good stability for mixing. The molecular weight distribution of the glutenin polymers is an important factor; a shift toward higher molecular weight (larger polymers on the average) leads to a stronger, more elastic, more viscous dough. Certain glutenin subunits facilitate polymer formation by having two or more cysteine residues that participate in intermolecular disulfide bond formation, whereas others having only a single cysteine residue available for intermolecular disulfide bond formation interfere with formation of larger polymers.

In addition to intrinsic genetic factors in any given wheat cultivar, there are strong environmental effects (soil conditions, weather conditions) on breadmaking qualtiy. The molecular weight distribution of the glutenin polymers or the ratio of monomeric proteins to polymeric proteins in the grain seems likely to play a role in variations that arise from environmental effects. Changes in these characteristics may result from changes in the degree of expression of specific protein components, even though the same qualitative complement is present in a given cultivar. A good balance of dough elasticity versus dough extensibility is important to the wheat processors, and this balance differs for different types of products.

Celiac disease. In some individuals, various gluten proteins and the albumin-globulin proteins of wheat endsoperm elicit allergic reactions, such as respiratory distress or skin wheals. However, wheat and its close relatives, rye and barley, are notable for their involvement in celiac disease (gluten-sensitive enteropathy). Celiac disease is also likely to involve an immune response, but one that appears to differ from allergic reactions involving primarily IgE-type immunoglobulins. In susceptible individuals, gluten proteins and their counterparts in rye and barley produce changes in the small intestinal mucosa that, in more severe cases, damage the absorptive epithelium. Because almost all nutrients—proteins, carbohydrates, calcium, iron, vitamins, and so forth—are absorbed by way of the epithelium, there can be a myriad of wide-ranging symptoms, from diarrhea and loss of weight to osteoporosis and anemia.

Although the exact mechanism that initiates processes leading to mucosal damage in celiac disease is unknown, a current hypothesis involves the binding of peptides derived from gluten to major histocompatibility complex class II proteins, which in turn present these peptides to T cells. Damage may result from cytokine release as a consequence of T-cell activation. All gliadin proteins, and probably some of the glutenin subunits, have the activating factor, which corresponds to particular combinations of amino acids in the peptide sequences and almost certainly includes glutamine and proline residues. Although gliadins have 260 or more amino acid residues in their polypeptide chains, synthetic peptides having sequences corresponding to as few as 12 amino acids of the α-type gliadin sequence have been reported to be active. However, the unifying structural principle of active peptides remains to be determined. *See* ALLERGY; GASTROINTESTINAL TRACT DISORDERS; IMMUNITY.

Donald D. Kasarda

Processing

Milling has evolved from rudimentary crushing or cracking to sophisticated separation and refining. The main purpose of milling is isolation of the starch-protein matrix, that is, separation of the endosperm from the high-fiber bran and high-lipid germ. Under optimal conditions, milling yields a high-quality, uniformly colored flour with a relatively stable shelf-life. The flours of hard wheats (11 to 13% protein) develop strong gluten complexes during mixing and are therefore suitable for making bread. Whole soft wheats (9 to 11% protein) yield flours that are used primarily for cakes, cookies, and pastries. Durum wheat is used to produce a relatively coarse flour, semolina, used for manufacture of pasta products.

Cleaning. Freshly harvested grain may contain sticks, stones, soil, chaff, and other classes of mixed grains. Wheat must be clean and somewhat moist in order to be milled to a highly-quality flour. Even small amounts of extraneous materials can produce wide variations in the acceptability and subsequent baking performance of wheat flour.

Middling streams are segregated and directed to reduction rolls, which have smooth surfaces and turn at equal speeds. They further reduce the size of relatively pure chunks of endosperm. After each reduction roll, the stream is sized. Particles that are the proper size are removed as finished flour; oversized particles are scalped from the screens and returned to a reduction stream.

The term flour extraction describes the proportion (by weight) of flour obtained from a given weight of wheat and, therefore, the degree to which endosperm has been separated or refined from the whole kernel. The higher the flour extraction, the less bran has been removed from the flour. The total of all individual flour streams constitutes straight flour, generally 72% of the wheat kernel. Individual flour streams comprising straight flour may be blended to produce the various commercial flour grades.

Developments in milling technology have enabled single-pass regrinding of flours in a pin mill to reduce

particle size to ranges from 1 to 50 micrometers. This regrinding process followed by air classification of particles by size and density enables partitioning of protein (fine) and starch (coarse) fractions. In this manner, air-classified flours may be established with specified levels of protein or starch, each with designated functional properties. *See* FOOD MANUFACTURING.

Mark A. Uebersax

Wheat Biotechnology

The application of biotechnological tools such as deoxyribonucleic acid (DNA) sequencing and plant transformation has had a major impact on the understanding of the contribution of the wheat endosperm prolamine proteins, particularly the high-molecular-weight (HMW) glutenins, to wheat protein and dough chemistry.

DNA sequencing. The isolation and DNA sequencing of the HMW-glutenin genes has enabled scientists to completely understand the primary structure of these proteins, which was extremely difficult to determine using just the proteins themselves because of the large repetitive regions in their amino acid sequences. These primary protein amino acid sequences revealed the two main features likely to be responsible for the unique characteristics of the HMW glutenins: (1) the long, central, repetitive domains theorized to contribute some of the elastic properties of doughs, and (2) the placement in the two ends of each polypeptide of most disulfide bond-forming cysteine amino acid residues necessary for glutenin polymer formation. These genes have also been used to express the proteins in microbial systems, a technique that allows easier isolation and study of individual natural and modified HMW glutenins—studies not feasible using whole wheat seeds with large numbers of chemically similar proteins. Synthesis of these wheat recombinant HMW glutenins are resulting in new understandings of these critical proteins. *See* BIOTECHNOLOGY; DEOXYRIBONUCLEIC ACID (DNA).

Plant transformation. The HMW-glutenin genes were also the first genes to be used in experiments in bioengineering wheat, in attempts to specifically alter an agronomic or economically important trait. In the first experiments, natural or modified HMW-glutenin genes were inserted into wheat cells, and whole fertile plants were regenerated (transgenic plants). When transgenic seeds of later generations were tested for dough physical properties, it was found that addition of these HMW-glutenin genes led to the synthesis of additional proteins in the wheat seed. These transformation experiments indicate that it is possible to obtain stable, high-level expression of HMW-glutenin transgenes in wheat, with dramatic changes in the physical properties of the resulting dough. However, this needs to be confirmed by large-scale field testing of these transgenic wheat lines under different environmental conditions. However, it is now that in the next few years transgenic wheat lines will make major contributions to the understanding of the function of the glutenin proteins. In addition, by further understanding the molecular basis of the glutenin contribution to dough properties, it will be possible to design new wheat varieties for targeted end uses. *See* GENETIC ENGINEERING.

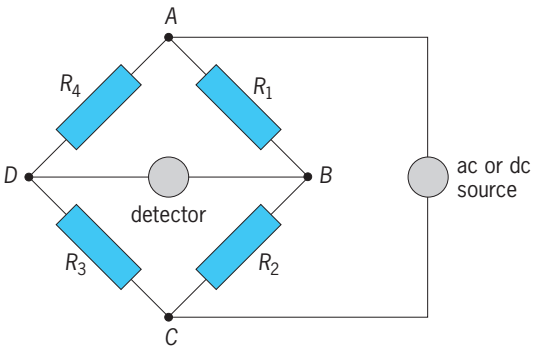
One advantage of modern biotechnological techniques is that genetic traits can be moved among plants with an ease and degree never before possible. For example, it is expected that the addition of the wheat HMW glutenins to other seeds, such as barley or maize, will alter the processing characteristics of flours from those crops and potentially lead to novel products and new markets. Although such biotechnological tools are extremely powerful and will become even more important in the study of wheat quality, these tools must still be coupled with studies of protein chemistry and dough functionality to relate genetic changes to the phenotypic traits important for wheat utilization.

Olin D. Anderson

Bibliography. Y. P. Bajaj, *Wheat*, 1991; F. Barro et al., *Nat. Biotech.*, 15:1295–1299, 1997; A. E. Blechl and O. D. Anderson, Expression of a novel high-molecular-weight glutenin subunit gene in transgenic wheat, *Nat. Biotech.*, 14:875–879, 1996; L. T. Evans and W. J. Peacock (eds.), *Wheat Science: Today and Tomorrow*, 1981; D. D. Kasarda, Gluten and gliadin: Precipitating factors in coeliac disease, in M. Maki, P. Collin, and J. K. Visakorpi (eds.), *Proceedings of the 7th International Symposium on Coeliac Disease* pp. 195–212, Coeliac Disease Study Group, Institute of Medical Technology, University of Tampere, Finland; J. H. Klippart, *The Wheat Plant: Its Origin*, *Culture*, *Growth*, *Development*, *Composition*, *Varieties Together with Information on Corn Culture*, 2 vols., 1980; W. Q. Loegering, J. W. Hendrix, and L. E. Browder, *The Rust Diseases of Wheat*, USDA Agr. Handb. 334, November 1967; M. Mäki and P. Collin, Coeliac disease, *Lancet*, 349:1755–1759, 1997; S. A. Matz, *Cereal Technology*, 1970; Y. Pomeranz, *Wheat*, *Chemistry and Technology*, 2 vols., 3d ed., 1988; E. J. Pyler, *Baking Science and Technology*, 2 vols., 1988; L. P. Reitz, *Wheat in the United States*, USDA Agr. Inform. Bull, 386, 1976; P. R. Shewry, N. G. Halford, and A. S. Tatham, The high molecular weight subunits of wheat glutenin, *J. Cereal Sci.*, 15:105–120, 1992; P. R. Shewry, M. J. Miles, and A. S. Tatham, The prolamin storage proteins of wheat and related cereals, *Prog. Biophys. Mol. Biol.*, 61:37–59, 1994; N. Shani, et al., Role of the amino-and carboxy-terminal regions in the folding and oligomerization of wheat high molecular weight glutenin subunits, *Plant Physiol.*, 98:433–441, 1991; R. Tkachuk and W. Bushuk, *Gluten Proteins*, 1990; W. T. Yamazaki, *Soft Wheat: Production*, *Breeding*, *Milling*, *and Uses*, 1981.

Wheatstone bridge

An extensively used electrical network of four resistances R_{1-4} (or impedances Z_{1-4} if alternating currents are involved). This network (see **illustration**), was first described by S. H. Christie in 1833, only 7 years after Georg S. Ohm discovered the relation-



Wheatstone bridge circuit.

ship between voltage and current. Since 1843, when Charles Wheatstone called attention to Christie's work, Wheatstone's name has been associated with this network. *See* ELECTRICAL IMPEDANCE; ELECTRICAL RESISTANCE; OHM'S LAW.

The potentials at nodes B and D with respect to node C (or A), depend only on the ratios R_1/R_2 and R_4/R_3 respectively, and some or all of the values of $R_{1\text{-}4}$ can be adjusted until these potentials are equal. A detector connected between B and D will then indicate zero current or voltage and Eq. (1) will be satisfied [or Eq. (2) for alternating-current impedances].

$$R_1/R_2 = R_4/R_3 \tag{1}$$

$$Z_1/Z_2 = Z_4/Z_3 \tag{2}$$

The network possesses the following useful properties:

1. Obtaining the detector indication of zero current or voltage (which is often termed "balancing the bridge") and the consequent validity of Eq. (1) or Eq. (2) is independent of the internal resistance (or impedance) of the source or detector.

2. The same result is obtained if the positions of source and detector in the network are interchanged. This is a simple example of reciprocity in linear systems. *See* RECIPROCITY PRINCIPLE.

3. From Eq. (1) or Eq. (2), the value of one of the four resistances or impedances can be calculated if one of the others and the ratio of the remaining two are known.

4. The balanced network will eliminate unwanted electrical signals (interference and pickup) between the source and circuitry connected to B and D.

The response of the detector to a deviation from the condition of Eq. (1) or Eq. (2) depends on the values of $R_{1\text{-}4}$ or $Z_{1\text{-}4}$ as well as the internal resistances or impedances of both source and detector. This response can be calculated by network analysis. *See* NETWORK THEORY; RESISTANCE MEASUREMENT.

Bryan P. Kibble

Wheel and axle

A wheel and its axle or, more generally, two wheels with different diameters or a wheel and drum, as in a windlass, rigidly connected together so that they rotate as a unit on a common axis. The principle of operation is the same as that of the lever in that, for static equilibrium, the summation of torques about the axis of rotation equals zero. Where flexible members, such as ropes, have been firmly attached to a wheel and drum (see **illus.**) and the machine is mounted on frictionless bearings, Eq. (1) applies.

$$F_1R_1 - F_2R_2 = 0 \tag{1}$$

If F_1 is the input and F_2 the output, the mechanical advantage is given by Eq. (2).

$$MA = \frac{F_2}{F_1} = \frac{R_1}{R_2} \tag{2}$$

If the bearings are not frictionless, a friction torque T_f will oppose the rotation. For the case with F_1 the input, the friction torque will act in the same direction as the load or output torque F_2R_2, and the equilibrium equation takes the form of Eq. (3), which may be solved for individual components as in Eq. (4)

$$F_1R_1 - F_2R_2 - T_f = 0 \tag{3}$$

$$F_2 = \frac{F_1R_1 - T_f}{R_2} \tag{4}$$

See ANTIFRICTION BEARING; TORQUE.

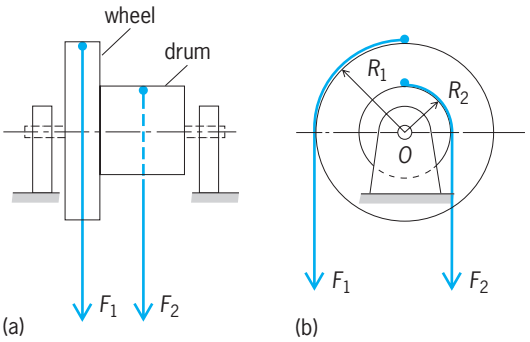
Because all members are rigidly connected, all torques act through the same angular displacement; the efficiency η is then given by Eq. (5).

$$\eta = \frac{F_2R_2}{F_1R_1} \times 100 = \frac{F_1R_1 - T_f}{F_1R_1} \times 100$$

$$= \left(1 - \frac{T_f}{F_1R_1}\right) 100 \tag{5}$$

The main difference between the lever and the wheel and axle is that the wheel and axle permits the forces to operate through a much greater distance. In the illustration the wheel and drum could be allowed to rotate any number of revolutions if the ropes were wrapped the required number of times around each before they were attached. *See* LEVER.

The wheel and axle is seldom encountered in the exact form shown in the illustration, but the principle of the torques about the axis of rotation



Wheel and axle. (*a*) Side view. (*b*) Front view.

equaling zero finds widespread applications in machines such as the belt drive. *See* FORCE; SIMPLE MACHINE.

Richard M. Phelan

Wheelbase

The distance in the direction of travel from front to rear wheels of a vehicle, measured between centers of ground contact under each wheel. For a vehicle with two rear axles, the rear measuring point is on the ground midway between rear axles. Tread of a vehicle is the distance perpendicular to the direction of travel between front wheels, or between rear wheels, measured from centers of ground contact.

Frank H. Rockett

White dwarf star

The smallest kind of ordinary star, about the size of the Earth. In most white dwarf stars, an amount of matter equal to 60% of the Sun's mass is compressed into this small volume. One quarter liter (about 1 cup) of white dwarf material has a mass of 600 tons. Several thousand white dwarf stars have been discovered.

Some of the other properties of white dwarf stars reach extreme values. The hottest known stars are either white dwarfs or stars that are just about to become white dwarfs, and have temperatures of a few hundred thousand kelvins. A few white dwarf stars have very strong magnetic fields, with field strengths exceeding 10^8 gauss (10^4 teslas), many times stronger than can be generated in laboratories. Some white dwarf stars rotate every few minutes, and some others rotate so slowly that no spin has been detected over several decades. The coolest white dwarf stars emit less energy per second than any other type of visible star. *See* HIGH MAGNETIC FIELDS.

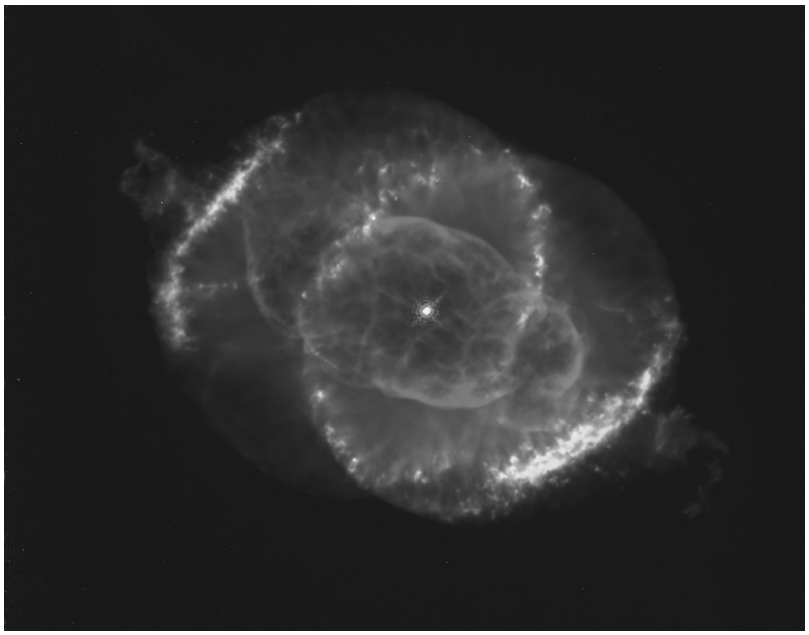


Fig. 1. Hubble Space Telescope image of the planetary nebula NGC 6543, nicknamed the Cat's Eye Nebula, surrounding a very hot, tiny blue star which is about to become a white dwarf star. (*P. Harrington, NASA*)

Astrophysical context. While these extreme stellar properties give white dwarfs some intrinsic interest, they fit into an astrophysical context that connects them to the universe at large. They are the final stages in the life cycles of low-mass stars like the Sun, the most common types of stars. At present, the Sun is on the main sequence, fusing hydrogen to form helium in its core. Five billion years from now, the hydrogen in the center of the Sun will run out, and the Sun will become a red giant star as it turns to other nuclear reactions to provide its internal heat. Stars like the Sun will become only hot enough inside that they will fuse helium nuclei to form a mixture of carbon and oxygen. They will then reach the end of the nuclear fusion road, and will no longer be able to generate their own sources of energy. At this point in their life cycle, they will be on the way to becoming white dwarf stars, the most common types of stellar remnants. *See* NUCLEOSYNTHESIS; STELLAR EVOLUTION.

Just before a star becomes a white dwarf star, it will shed its outer layers as its stellar wind strengthens at the very end of the red giant stage of its life cycle. In many and perhaps all cases, the expelled gas will become visible as a glowing gas cloud called a planetary nebula. The confusing name of these objects was given by nineteenth-century astronomers, who knew nothing of their origin but surmised only that these roundish gas clouds looked somewhat like Uranus and Neptune. **Figure 1** shows an interesting planetary nebula: the dot in the middle of the swirls of gas is a very hot, blue star which is about to become a white dwarf. *See* PLANETARY NEBULA.

These connections to the prior phases of stellar life cycles relate white dwarf stars to the rest of astrophysics. For example, the formation of white dwarf stars changes the chemical composition of the Milky Way Galaxy as a whole. When stars eject mass at the end of their lives, this ejected mass enriches the interstellar medium, the gas between the stars. This mass will eventually condense to become second- and third-generation stars and planets. The ejected mass is enriched in elements such as carbon, oxygen, and nitrogen, which are critical for life and human civilization. *See* INTERSTELLAR MATTER.

Structure. A star becomes a bona fide white dwarf when it contracts to a state where heat pressure plays a very small role in determining its internal structure. When the first white dwarfs were discovered in the 1920s, senior astronomers such as Arthur Eddington and Henry Norris Russell were astonished because it seemed impossible that stellar material could be compressed to such enormously high densities. Subrahmanyan Chandrasekhar, traveling from India to England to begin his graduate studies in astronomy, first recognized that the material in white dwarf stars was a previously unrecognized form of matter, called degenerate matter. The electron clouds in degenerate matter actually touch each other, and this touching gives white dwarf matter a resistance

to compression. This resistance to compression balances the force of gravity.

The initial reaction to these ideas of Chandrasekhar was one of disbelief. However, his theory of white dwarf stars can be tested by the use of the mass-radius relation. Since the heat in a star plays no role in its internal structure, a star of a particular mass can have only a particular radius, as long as it has the same interior chemical composition as that of all other white dwarf stars. Chandrasekhar's theoretical explanation predicts that all white dwarfs composed of a carbon/oxygen mix should fall along the same line in a graph which plots stellar mass against stellar radius. Later modifications to his theory suggest that some hot white dwarf stars should lie slightly above Chandrasekhar's line, since heat pressure plays a small role in their structure.

Tests of this theory began in the 1960s, but early data were fragmentary. The launching of the European *Hipparcos* spacecraft in 1989 provided astronomers with the distances to a large number of stars and permitted a more definitive test. The results are shown in **Fig. 2**, in which the general agreement between Chandrasekhar's theory and the observations is apparent. Several stars fall below the line for carbon, which the interior of white dwarf stars is expected to contain. Astronomers are still undecided as to whether some white dwarf stars have interiors made of iron or whether there are some observational uncertainties that need to be cleared up.

Life cycle. Superficially, the life cycle of white dwarf stars is simple. They have no nuclear energy sources, so they simply cool. The hottest star to have compacted itself to the white dwarf state is H1504+65, with a temperature of 190,000 K (350,000°F). The coolest white dwarf stars have temperatures near 3800 K (6400°F), still hotter than the coolest main-sequence stars. The internal structure of these two stars is virtually the same. Consequently, white dwarf stars are spread all over the bottom of the Hertzsprung-Russell diagram, a graphical tool which is used to classify stars. *See* HERTZSPRUNG-RUSSELL DIAGRAM.

The temperatures of the coolest white dwarf stars can be used to determine how long they have cooled. The numbers of very cool white dwarf stars drop off significantly when their temperatures reach approximately 5000 K (8500°F), about the temperature of the Sun. White dwarf stars which are this cool have been cooling for several billion years. More quantitative calculations of their lifetimes and allowance for the length of time it takes a star to become a white dwarf star can provide an estimate of the age of the stellar population of the Milky Way Galaxy. The derived age is 8–10 billion years. This age is marginally consistent with measurements of the ages of other rather old objects in the universe such as the globular star clusters, which are thought to be at least 12 billion years old.

However, the life cycle of white dwarf stars, when considered in a little more depth, is more complex. The surface layers of white dwarfs are much more varied than is the case of other types of stars. The most obvious variation is in the ratio of hydrogen to helium, the two most common elements in the universe. In virtually every other type of star, the hydrogen/helium ratio is near its cosmic value of approximately 10:1. However, white dwarf stars come in two basic flavors, hydrogen-rich and helium-rich. The hydrogen-rich stars have a surface composition which is very hydrogen-rich indeed, with no traces of helium. In some of the best-observed cases, the hydrogen/helium ratios exceed 10^7. In the helium-rich star, the surfaces are very much depleted in hydrogen. In the best-observed cases, the hydrogen/helium ratio is less than 10^{-7}. This variation in the surface hydrogen/helium ratio, over 14 orders of magnitude, is one of the most extraordinary chemical peculiarities in stellar astronomy. It seems likely that at least part of the variation comes from subtle

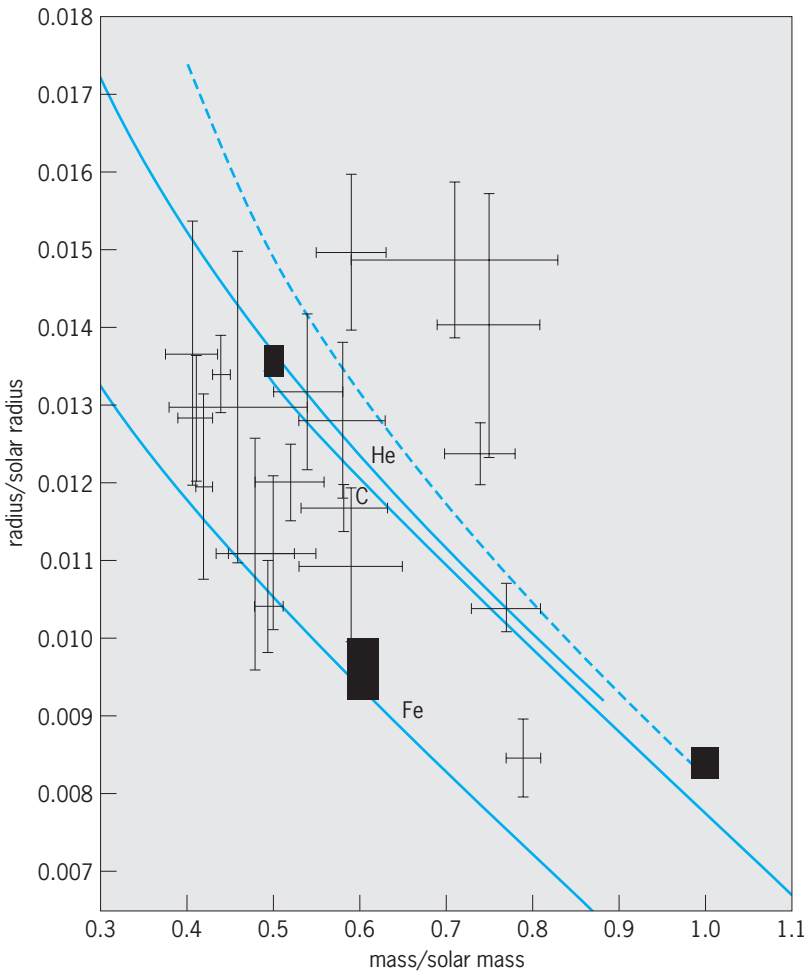


Fig. 2. Mass-radius relation for white dwarf stars. The solid lines give the relationship between mass and radius which is expected for white dwarf stars composed of helium (He), carbon (C) or a carbon/oxygen mixture, and iron (Fe). The broken line gives a revised version of the theory, which includes the hydrogen surface layer as well as the carbon interior. The data points and error bars give the position of a number of observed white dwarf stars. The mass and radius of each star are at the intersection of the vertical and horizontal lines. The three boxes are the positions of white dwarf stars in binaries, which are the measurements of highest precision and fewest assumptions. The agreement of these boxes with the theoretical lines supports Chandrasekhar's theory. (*After H. Shipman and J. Provencal, Confirming a Nobel Prize winning theory with the Hipparcos spacecraft: Accurately determining distances and diameters of white dwarf stars, in J.-E. Solheim and E. Meistas, eds., 11th European Workshop on White Dwarf Stars, pp. 15–23, Astronomical Society of the Pacific San Francisco, 1999*)

differences in the way that white dwarfs form in the late stages of red giant life cycles.

When white dwarf stars reach certain temperatures, sizes, and chemical compositions, their interiors become unstable and they vary in the amount of light that they emit. Waves of brighter and darker parts of the surface sweep around the star, producing the variations in their light from what are technically called nonradial pulsations. The periods of white dwarf stars range from 20 minutes to several hours. As a result, their pulsations are best observed by worldwide networks of telescopes, since any individual site can observe only a few pulsational cycles before the star sets. Slow changes in the temperature and diameter of one white dwarf star, PG1159–035, have been detected in the study of its pulsations. The time scale is roughly a million years, about what is expected from understanding of stellar evolution. Similar changes should be observable in other stars.

Perhaps the most extreme and least understood aspect of white dwarf stars is their high magnetic fields, which range from 1 megagauss (100 teslas) to a few hundred megagauss (a few tens of thousands of teslas). It is believed that these fields are fossil fields, the remnants of much weaker fields which exist in some stars that contain a strong, overall magnetic field. The fields in white dwarfs are discovered because they shift the wavelengths of spectral lines in the stars' surfaces. *See* STAR; ZEEMAN EFFECT. Harry L. Shipman

Bibliography. J. B. Kaler, *Stars and Their Spectra*, 1989; J.-E. Solheim and E. G. Meistas (eds.), *11th European Workshop on White Dwarfs*, 1999; V. Weidemann, Masses and evolutionary status of white dwarf stars and their progenitors, *Annu. Rev. Astron. Astrophys.*, 28:103–137, 1990.

Whooping cough

An acute infection of the tracheobronchial tree caused by *Bordetella pertussis*, a bacteria species exclusive to infected humans. The disease (also known as pertussis) follows a prolonged course beginning with a runny nose, and finally develops into violent coughing, followed by a slow period of recovery. The coughing stage can last 2–4 weeks, with a whooping sound created by an exhausted individual rapidly breathing in through a narrowed glottis after a series of wrenching coughs. The classical disease occurs in children 1–5 years of age, but in immunized populations infants are at greatest risk and adults with attenuated (and unrecognized) disease constitute a major source of transmission to others.

Bordetella pertussis is highly infectious, particularly following face-to-face contact with an individual who is coughing. The inhaled bacteria attach specifically to the hairlike cilia of cells lining the respiratory airway and remain there without invading host tissues. The disease is caused by structural components and extracellular toxins elaborated by *B. pertussis* from this superficial niche. Multiple virulence factors produced by the organism play important roles at various stages of pertussis.

For decades a vaccine produced from whole *B. pertussis* cells and combined with diphtheria and tetanus toxoids has been used throughout the world for routine childhood immunization. This vaccine has been effective in prevention of childhood pertussis, although immunity declines after 12 years. This whole-cell vaccine has significant side effects (pain, swelling, fever). Concern over vaccine morbidity has caused immunization rates to decline in some developed countries. These drops in immunization rates have often been followed by widespread outbreaks of disease, including deaths. Considerable effort has been directed toward the development of a vaccine which would minimize side effects but maintain efficacy through use of purified *B. pertussis* virulence factors. A new acellular vaccine containing filamentous hemagglutinin and pertussis toxin is available and has fewer side effects than the whole-cell vaccine. The acellular vaccine has been in use for some time in Sweden, Japan, and England. It has been approved for primary immunization in the United States. *See* DIPHTHERIA; TETANUS; VACCINATION.

Although *B. pertussis* is susceptible to many antibiotics, their use has little effect once the disease reaches the coughing stage. Erythromycin is effective in preventing spread to close contacts and in the early stage. *See* DIPHTHERIA. Kenneth J. Ryan

Bibliography. K. J. Ryan (ed.), *Sherris Medical Microbiology*, 3d ed., 1994.

Wide-area networks

Communication networks that are regional, nationwide, or worldwide in geographic area, with a minimum distance typical of that between major metropolitan areas. Smaller networks include metropolitan and local-area networks. A communication network provides common transmission, multiplexing, and switching functions that enable users to transport data between many sources and many destinations. Under ideal circumstances, the data that arrive at the destination are identical to the data that were sent. The rate of arrival of bits at any point in the network is said to be the data rate at that point and is typically measured in bits per second. These bits may come from one source or from a multiplicity of sources. The capacity of a network to transmit at a cerain data rate is known as its bandwidth. *See* LOCAL-AREA NETWORKS.

Network fundamentals. There are several fundamental attributes and concepts that facilitate the accurate transmission of data within and between digital networks. To communicate between computers, a set of rules, formats, and delivery procedures known as protocols must be established. Once these are agreed to, an Interface Control Design (ICD) is created as the basis for the communications between the sender and receiver of data.

Part of the communications protocol allows for definition of where the packets of digital data are to be routed. Each packet of data contains the unique

address of a computer or other network as its destination. The routing of the data is known as packet switching since the nodes in the network can switch the packet to various transmission paths. Networks allow alternate or redundant communications paths in the case of network transmission problems. Networks are interconnected by means of routers. These devices allow packets received from one network to be passed along to another network, while insulating each network from the internal characteristics of the other. Thus, two networks with different internal transmission speeds can be connected successfully by means of a router. *See* PACKET SWITCHING.

Another part of the communications protocol allows for including error detection and correction information in the data packets. The destination computer or network will verify the data in the packet utilizing the error control data, such as a checksum. If the data are found to be corrupted, a negative acknowledgment (NACK) will be returned to the sender, and the packet must be resent.

Protocols are also used to implement flow control. This allows the receiving computer or network to communicate back to the sender when it can or cannot receive additional data. This is important in that both computers and networks vary in terms of processing speeds and bandwidth capacity.

Advances. Since the early 1990s, the ubiquity and deployment of wide-area networks (WANs) such as terrestrial, wireless, and satellite networks have exploded around the world. It is now possible for nearly any person or entity to interface with a network. Some of these entities include personal hand-held communications devices, home and office appliances (for example computers, telephones, televisions, and security monitors), automobiles and trucks, and land- or water-based infrastructures (such as highway systems, oil rigs, and irrigation systems). *See* DIGITAL COMPUTER.

Protocols. Certain protocols have become standards for a majority of wide-area networks. Asynchronous Transfer Mode (ATM) is a protocol used in business-to-business (B2B) communications when high data rates are needed. Typically, one ATM port supports 45 megabits per second (Mbps). Frame Relay is another business-to-business protocol. The advantage of Frame Relay is that the data rate can be scaled to the individual company's needs. As an example, one company may purchase a 512 kilobits per second (kbps) frame relay circuit, whereas a company with lower requirements may purchase a 256-kbps circuit. These two protocols can also coexist in a hybrid network by utilizing devices known as protocol converters. Companies use this flexibility to place voice, video, and data simultaneously on the same circuits using a packet switching methodology. The third protocol, and the one having the most worldwide impact on both business and personal communications, is the Internet Protocol, or IP. *See* INTEGRATED SERVICES DIGITAL NETWORK (ISDN).

Transmission media and architecture. Wide-area networks may operate on a mix of transmission media for either fixed or mobile applications. Fixed applications mean that the receiver of digital data is stationary. Examples of wireline transmission media for fixed applications are fiber-optic cable, copper wire, and coaxial cable. Many advances in the deployment of wireline media have significantly increased their capacity for handling digital data. For copper, a technique known as digital subscriber line (DSL) allows for transmission in excess of 1 Mbps over regular phone lines. In fiber, a technique known as wave division multiplexing (WDM) allows the simultaneous transmission of different streams of digital data over each spectral component of the light wave. This allows bundles of fiber-optic cable to transport billions of bits (gigabits) and even trillions of bits (terabits) of data per second. Advances in software and hardware-based digital compression also contribute to the ability to transmit more information on a specific medium over a set amount of time. *See* COMMUNICATIONS CABLE; OPTICAL COMMUNICATIONS; OPTICAL FIBERS; TRANSMISSION LINES.

Wide-area networks also operate over a variety of wireless media. Wireless media can support either fixed or mobile applications. Another common distinction in wireless networks is whether it is point-to-point or point-to-multipoint. In point-to-point the originating transmission has one receiver, whereas in point-to-multipoint the originating transmission has multiple receivers. Examples of wireless media include radio-wave, microwave, cellular, and satellite. Different wireless media operate over different parts of the radio spectrum. The allocation and control of this spectrum is managed by the Federal Communications Commission (FCC). *See* COMMUNICATIONS SATELLITE; MOBILE RADIO; RADIO SPECTRUM ALLOCATIONS; RADIO-WAVE PROPAGATION.

Content. There are many different types of content transmitted over WANs. Examples of content are data, voice, video, audio, paging messages, and fax. By virtue of the ability to digitize all of this content, the major difference in transmission requirements is the bandwidth, or capacity, required to transmit digital packets of any type of content. For instance, it takes much more bandwidth to transmit digital video (1-megabit range) than it does to transmit digital audio (64-kilobit range). Certain types of content, such as video and audio, require a network and protocol that deliver the content in real time. The data packets must arrive at the destination in a particular sequence. Other types of content, such as financial data, do not have this requirement. To accommodate these differences, a specific type of WAN architecture, in terms of speed and capacity, must be designed and sized for the content application. *See* DATA COMMUNICATIONS; ELECTRICAL COMMUNICATIONS; FACSIMILE.

Digital compression. In order to utilize the bandwidth capacity of WANs more efficiently, digital compression techniques are now used in many applications. Fundamentally, digital compression reduces the amount of bits in data packets by removing repetitive strings of bits and replacing them with shorter packets that numerically describe the amount of repetitive data. An example is a television picture

with a preponderance of blue sky. Using digital compression, this part of the picture can be represented with a relatively small amount of data. At the receiver site, data are decompressed using the information in the compression packets. The original content can then be almost fully reconstructed for display or processing. Very powerful software compression techniques have been implemented in a wide variety of user applications and consumer devices, such as digital cable and satellite television. This has been made possible by tremendous advances in microprocessor computing power and chipset memory capacities, allowing the implementation of very complex compression algorithms (or software programs) to squeeze more data into existing bandwidth using economical integrated circuit chips for decompression. *See* DATA COMPRESSION.

Digital data security. With massive amounts of information being transmitted over WANs using both public and private infrastructure, data security is increasingly important. In order to secure digital data transmitted over networks, encryption techniques have been developed. Encryption of digital data involves using hardware and software to manipulate the bits in a data packet, making it unrecognizable and unusable to anyone not authorized to use the data. An authorized receiving device on the network contains the corresponding decryption hardware and software to restore the data content to its original format. Often in digital networks, the data are both compressed and encrypted in the same transmission process. Due to the value of the information on the network, it may come under attack by unauthorized parties attempting to decrypt the data. To combat this, some networks are employing various forms of "renewable security," which can be updated from time to time. An example is the use of smartcards in receiving devices, which can be replaced when an encryption algorithm is upgraded. *See* COMMUNICATIONS SCRAMBLING; COMPUTER SECURITY; CRYPTOGRAPHY.

Deployment costs. In deploying WAN infrastructure, there are cost and performance trade-offs involved in determining the appropriate topology of the network. For wireline-based networks, the wire must be either buried underground or strung on existing poles. For wireless-based networks, transmitters must be installed on towers or buildings, or communications satellites must be launched into orbit. Sometimes the business model may dictate that a hybrid of several of these topologies be used. There are several satellite orbital topologies used for different networks: geosynchronous satellites (GEO) for satellite television, medium Earth orbit (MEO) and low Earth orbit (LEO) for telephone and Internet applications. A specific deployment cost factor encountered is known as the "last mile" phenomenon. This refers to the costs to deploy the network to the farthest geographic region of desired coverage. For instance, wireline topographics that are cost-effective in urban areas often cannot be justified in some suburban or rural areas. In this case, use of a hybrid wireless transmission or satellite coverage could be used to fill in these hard-to-access areas. The cost of deploying to these areas must be justified by either the additional business benefit or expected subscriber revenue. Another instance is the type of connection between the wideband transmission medium (such as fiber optics) and the subscriber homes it passes.

Internet. The Internet, using the IP, has become by far the most ubiquitous WAN in the world. Bandwidth and computing power have progressed to the point that Internet users are able to transmit almost any type of content over the network at higher speeds and reasonably good quality. Everything from video clips, electronic mail (e-mail), telephone calls (called Voice Over IP), to digitized x-rays can be transmitted over the Internet. There are three basic variations of IP networks. First is the overall Internet itself which encompasses all personal and business users of the Internet. The second type of IP-based networks is intranets. An intranet is usually deployed within a specific organization or company. A company intranet may be used to manage human resources and financial processes and keep employees updated on company news. The third type of IP-based networks is extranets. Typically, extranets are used to link multiple organizations or companies for some common business purpose. For example, an extranet may be created between an automaker and its parts suppliers to facilitate the processes of ordering, invoicing, and inventory control between the companies. In both intranets and extranets, a technology known as a firewall is employed to prevent unauthorized access to the network. Firewalls are a combination of hardware and software that filter out any unauthorized URL (Uniform Resources Locator) from the network. The URL is the basic unique address or location for any Web site or other Internet service. *See* ELECTRONIC MAIL; INTERNET.

Richard L. Jensen

Bibliography. H.C. Berkowitz, *WAN Survival Guide: Strategies for VPNs and Multiservice Networks*, 2000; R. Cahn and R. Chan, *Wide Area Network Design: Concepts and Tools for Optimization*, 1998; S. Feit, *Wide Area High Speed Networks*, 1999; J. S. Marcus, *Designing Wide Area Networks and Internetworks: A Practical Guide*, 1999; R. W. McCarty (ed.) and M. McGregor, *Cisco WAN Quick Start*, 2000; M. McGregor, *Cisco IOS 12.0 Wide Area Networking Solutions*, 1999.

Wiedemann-Franz law

An empirical law of physics which states that the ratio of the thermal conductivity of a metal to its electrical conductivity is a constant times the absolute temperature, as given by Eq. (1). Here K_c is the ther-

$$K_c = L_0 \sigma T \tag{1}$$

mal conductivity due to the conduction electrons, σ is the electrical conductivity, T is the absolute temperature, and L_0 is known as the Lorentz number. For the case of a degenerate electron gas, the value of L_0

is given by Eq. (2), where k is the Boltzmann constant and e is the electronic charge. The Wiedemann-Franz law provides an important check on theories of electrical and thermal conductivity. *See* CONDUCTION (HEAT); ELECTRICAL CONDUCTIVITY OF METALS; FREE-ELECTRON THEORY OF METALS; THERMAL CONDUCTION IN SOLIDS.

$$L_0 = \frac{\pi^2}{3}\left(\frac{k}{e}\right)^2 \qquad (2)$$

Frank J. Blatt

Bibliography. H. S. Carslaw and J. C. Jaeger, *Conduction of Heat in Solids*, 2d ed., 1986; C. Kittel, *Introduction to Solid-State Physics.*, 7th ed., 1995; J. M. Ziman, *Principles of the Theory of Solids*, 2d ed., 1979.

Wilkinson Microwave Anisotropy Probe

A spacecraft which is precisely measuring the cosmic microwave background (CMB) radiation. The *Wilkinson Microwave Anisotropy Probe* (*WMAP*) is a space mission of the National Aeronautics and Space Administration (NASA) that has put fundamental theories of the nature of the universe to a precise test. Since August 2001, *WMAP* has continually surveyed the full sky, mapping out tiny differences in the temperature of the cosmic microwave background radiation, which is the radiant heat from the big bang. A fossil remnant of the hot big bang, the cosmic microwave background permeates the universe and is seen today with an average temperature of only 2.725 kelvins. Tiny variations about this average temperature were first discovered by NASA's *Cosmic Background Explorer* (*COBE*) mission. *WMAP* followed up on the *COBE* results by characterizing the detailed statistical nature of the temperature variations of the cosmic microwave background (called anisotropy), revealing a wealth of detail about the global properties of the universe.

History. In developing the quantitative details of the big bang theory in the late 1940s, Ralph Alpher noted that a remnant radiation should still exist, cooled by the expansion of the universe from its hot and dense earliest stages. In 1965, the cosmic microwave background was discovered accidentally by Arno Penzias and Robert Wilson. In 1992, NASA launched the *COBE* mission, which measured the precise spectrum of the cosmic microwave background and detected its temperature variations at large angular scales (greater than 7°). Recognizing that a wealth of information about the physical conditions of the early universe is encoded in the statistics of the variations at finer angular scales, scientists proposed *WMAP* to map the cosmic microwave background with higher resolution (and greater sensitivity) than *COBE*. *See* BIG BANG THEORY.

WMAP is one of a series of NASA medium-class explorer (MIDEX) satellites. The satellite was launched by a Delta 2 rocket on June 30, 2001, at 3:46 p.m. EDT (1946 UTC) from the Cape Canaveral Air Force Station in Florida. Originally called the *Microwave Anisotropy Probe* (*MAP*), the satellite was renamed in 2003 to honor the memory and accomplishments of David T. Wilkinson, a member of the science team and a pioneer in cosmic microwave background studies.

The mission was proposed to NASA in 1995, selected in April 1996, and confirmed for development in 1997. The satellite, with its single instrument, was built, tested, and launched in only 4 years.

Control of systematic measurement errors. The *WMAP* science requirements dictated that the relative cosmic microwave background temperature be measured accurately over the full sky. The overriding design requirement was to control systematic errors that would otherwise contaminate the measurements. To achieve this, *WMAP* uses differential microwave radiometers that measure temperature differences between pairs of spots on the sky.

To facilitate the separation of the cosmic microwave background from foreground signals from the Milky Way Galaxy, *WMAP* uses polarization-sensitive radiometers at five separate frequency bands centered at 23, 33, 41, 61, and 94 GHz (wavelengths of 13, 9.1, 7.3, 4.9, and 3.2 mm). There are 1, 1, 2, 2, and 4 independent pairs of feed horns per frequency, respectively, with beam sizes of 0.88, 0.66, 0.51, 0.35, and 0.22°. The radiometers are rapidly modulated with a 2.5-kHz phase switch. Amplitude calibration (that is, the determination of the temperature scale of the instrument output signal) relies on the in-flight modulation of the cosmic microwave background temperature caused by the Doppler shift arising from the satellite's motion. Calibration of the beam pattern and angular resolution relies on in-flight observations of Jupiter.

Dual back-to-back Gregorian primary reflectors with apertures of 1.4 × 1.6 m (4.6 × 5.2 ft) focus the microwave radiation from two spots on the sky roughly 140° apart and feed these signals to 10 separate differential receivers that are in an assembly directly underneath the primary optics (**Fig. 1**). Large radiators between the primary optics passively cool the sensitive amplifiers in the receiver assembly below 90 K (−290°F). The bottom half of the spacecraft provides the necessary avionics functions, such as command and data collection electronics, attitude (pointing) control and determination, power services, and a propulsion system. The entire observatory is kept in continuous shadow by a large deployed sunshield, which includes the solar panels that power *WMAP*. *See* ANTENNA (ELECTROMAGNETISM).

Orbit. *WMAP* observes the sky from an orbit about the second Sun–Earth Lagrangian point (L2), 1.5×10^6 km (0.9×10^6 mi) from Earth (**Fig. 2***a*). L2 is about four times farther from Earth than the Moon. This vantage point offers an exceptionally stable environment since the observatory can always point away from the Sun, Earth, and Moon while maintaining an unobstructed view of deep space. *See* CELESTIAL MECHANICS.

Scanning the full sky is an important part of

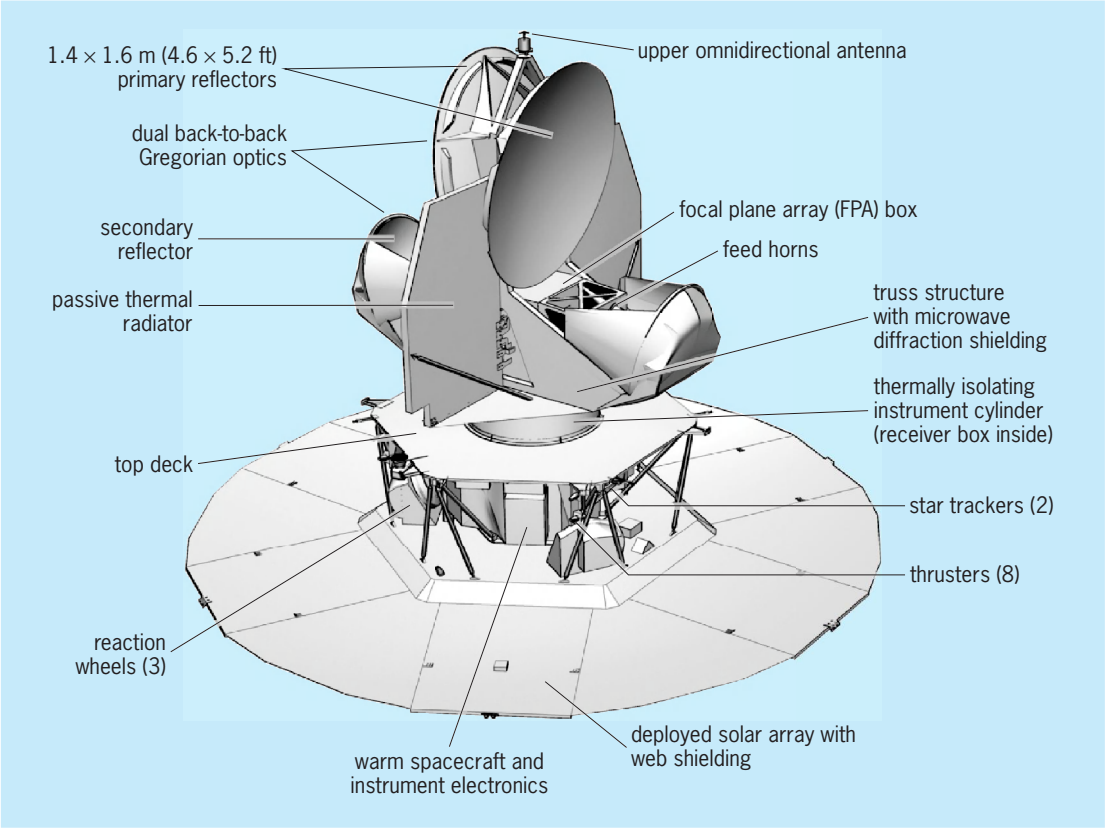


Fig. 1. ***Wilkinson Microwave Anisotropy Probe (WMAP)*** **spacecraft. (*NASA/WMAP Science Team*)**

WMAP's mission. *WMAP* scans the sky in such a way as to cover about 30% of the sky each day (Fig. 2*b*); as the L2 point follows the Earth around the Sun, *WMAP* observes the full sky every 6 months.

The attitude is controlled using two star trackers, two gyroscopes, coarse and fine sun sensors, and three reaction wheels. The satellite spins at 0.464 revolution per minute (~2 min per spin) and precesses at 0.017 revolution per minute (1 h per precession) about a 22.5° cone on the WMAP–Sun

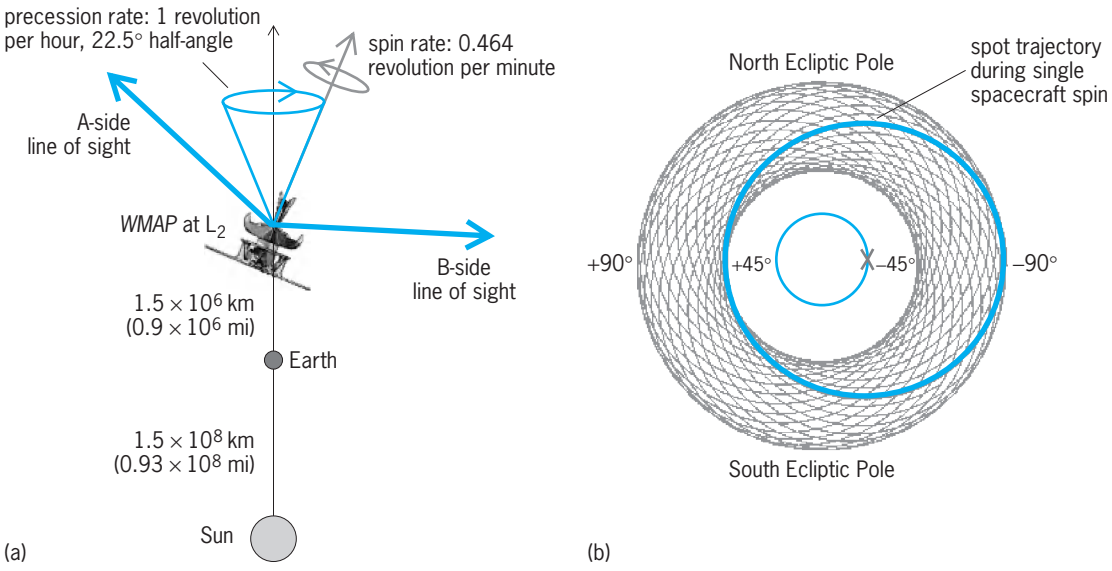


Fig. 2. ***WMAP*** **sky-scan strategy. (*a*) Location of the satellite relative to the Earth and Sun (not to scale), and the orientation of the satellite and of the lines of sight to the spots on the sky focused by the primary reflectors, which continually change as the satellite spins and precesses. (*b*) Sky pattern showing the trajectory of one of the spots focused by the primary reflectors. The ecliptic is horizontal and passes through the center of the diagram. Celestial longitude relative to the antisolar point is indicated. The precession of the spin creates the doughnut pattern that moves along the ecliptic in the course of a year. Full-sky coverage is achieved every six months. (*NASA/WMAP Science Team*)**

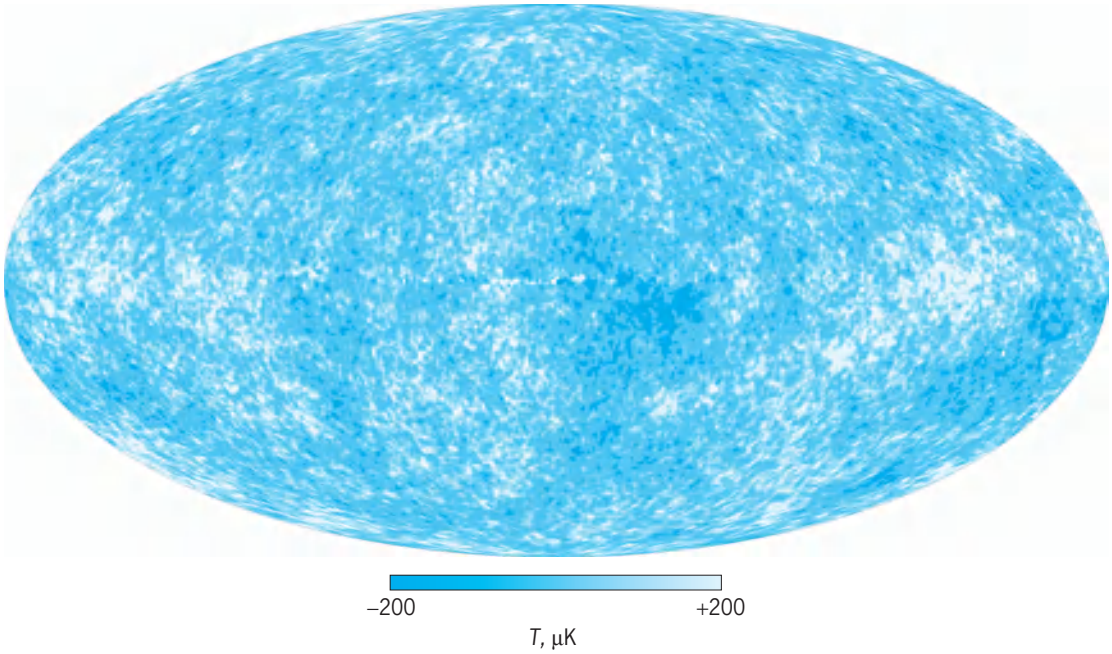


Fig. 3. ***WMAP*** **full-sky cosmic microwave anisotropy map. The microwave light captured in this picture from *WMAP* is from 379,000 years after the big bang, over 13 × 10^9 years ago. It is a "baby picture" of the universe, equivalent to taking a picture of an 80-year-old person on the day of his or her birth. Shades indicate "warmer" (light) and "cooler" (dark) spots; scale bar indicates variation (*T*) from average temperature. The oval shape is a projection of the full sky in galactic coordinates. This map is a composite from the five *WMAP* frequencies, and the effects of the Earth's proper motion are removed. (*NASA/WMAP Science Team*)**

line. Blowdown hydrazine propulsion with eight thrusters were used to achieve the L2 orbit, and are also used for station-keeping. *See* GYROSCOPE; STAR TRACKER.

Spacecraft design. The spacecraft structure is made of carbon composite and aluminum materials. The total observatory mass at launch was 840 kg (1850 lb). Communications employ two omnidirectional antennas and a fixed medium-gain antenna, which is used at 667 kilobits per second for daily downlinks to the 70-m (230-ft) antennas of the Deep Space Network. A 3.1-m^2 (33-ft^2) gallium arsenide germanium (GaAs/Ge) solar array oriented 22.5° off the full sun line, and a 23 A-h nickel-hydrogen (NiH) battery provide the required 419 W of power. There are no eclipses in the observing phase of the mission. The design lifetime of 27 months has been exceeded. *See* SPACECRAFT GROUND INSTRUMENTATION.

First results. The first results from the *WMAP* mission were reported on February 11, 2003. This release of the first year of flight data included the most detailed full-sky "baby picture" of the universe taken so far (**Fig. 3**). The *WMAP* data were compared and combined with other diverse cosmic measurements (based on galaxy clustering, supernovae, and so forth), and a new unified and precise understanding of the universe emerged.

The universe is 13.7 × 10^9 years old, with a margin of error of about 1%. The cosmic microwave signal in the *WMAP* map is from 379,000 years after the big bang. The expansion rate of the universe (the Hubble constant) value is $H_0 = 71$ (km/s)/Mpc, with a margin of error less than 5%. The contents of the universe include 4% baryons, 23% cold dark matter, and 73% dark energy. Fast-moving neutrinos do not play a major role in the evolution of structure in the universe. In fact, if they had played such a role, they would have prevented the early clumping of gas in the universe, delaying the emergence of the first stars, in conflict with the new *WMAP* data. *See* COSMOLOGY; DARK ENERGY; DARK MATTER; HUBBLE CONSTANT; NEUTRINO; UNIVERSE.

The *WMAP* data place new constraints on the dark energy. One possibility for the dark energy, Albert Einstein's cosmological constant, is fully consistent with the *WMAP* data. In this scenario the universe will expand forever. However, if the dark energy is due to some other physical effect, the conclusion about the fate of the universe could change. *See* COSMOLOGICAL CONSTANT.

The polarized data also provide new evidence for inflation, the rapid expansion of the universe a fraction of a second after its birth. Many particular inflation models are ruled out; others are supported with this new evidence. The inflationary theory predicts that density is very close to the critical density, producing a flat (Euclidean) universe. *WMAP* has determined, within the limits of instrument error, that the universe is flat. Thus, the results support the big bang and inflation theories. *See* COSMIC BACKGROUND RADIATION; INFLATIONARY UNIVERSE COSMOLOGY.

Charles L. Bennett

Bibliography. C. L. Bennett et al., First year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Preliminary maps and basic results, *Astrophys. J. Suppl.*, 148:1, 2003; C. L. Bennett et al., The Microwave Anisotropy Probe (MAP) mission, *Astrophys. J.*, 583:1–23, 2003; M. Chown, *Afterglow of Creation:*

From the Fireball to the Discovery of Cosmic Ripples, University Science Books, 1996; A. H. Guth, *The Inflationary Universe: The Quest for a New Theory of Cosmic Origins*, Perseus Books Group, 1998; M. D. Lemonick, *Echo of the Big Bang*, Princeton University Press, 2003; D. N. Spergel et al., First year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Determination of cosmological parameters, *Astrophys. J. Suppl.*, 148:175, 2003.

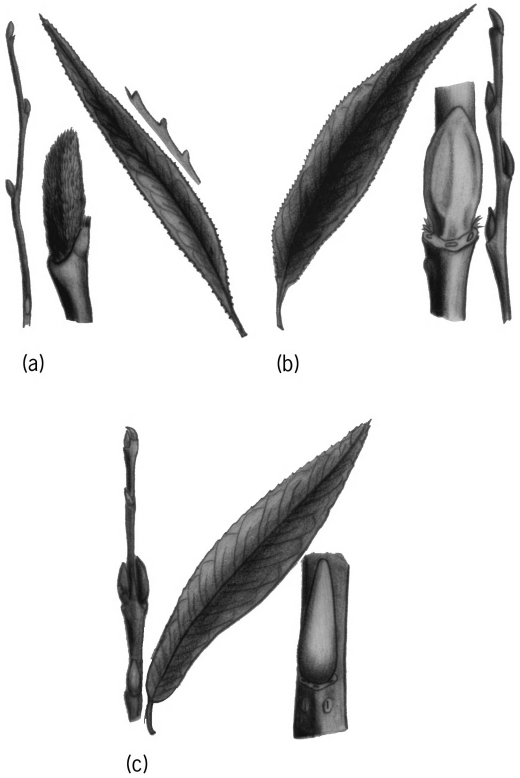
Willemite

A nesosilicate mineral, composition Zn_2SiO_4, crystallizing in the hexagonal system. It is usually massive or granular with a vitreous luster; crystals are rare. The mineral has a basal cleavage and may be variously colored, most commonly green, red, or brown. Hardness is $5\frac{1}{2}$ on Mohs scale; specific gravity is 3.9–4.2. Willemite forms a valuable ore of zinc at Franklin, New Jersey. At this famous zinc deposit willemite fluoresces a yellow-green and is found in crystalline limestone associated with franklinite, zincite, and many other rarer minerals. With the exception of the occurrence in New Jersey, willemite is a rare mineral. *See* SILICATE MINERALS.

Cornelius S. Hurlbut, Jr.

Willow

A deciduous tree and shrub of the genus *Salix*, order Salicales, common along streams and in wet places in the United States, Europe, and China. The twigs are often yellow-green and bear alternate leaves which are characteristically long, narrow, and pointed, usually with fine teeth along the margins (see **illus.**). Flowers occur in catkins; the staminate consist of one to twelve (usually two) stamens, with one or two nectar glands; the pistillate have a single ovary of two carpels, with one to four small glands. The fruit contains several silky seeds. The species are wind- or insect-pollinated and hybridize freely, which often makes identification difficult. *See* SALICALES.



Willow diagnostic features. (*a*) Babylon weeping willow (*Salix babylonia*), twig, terminal bud, and leaf. (*b*) Crack willow (*S. fragilis*), leaf, lateral bud, and twig. (*c*) Purple-osier willow (*S. purpurea*), twig, leaf, and lateral bud.

Willow lumber is used for fuel and in making charcoal, excelsior, ball bats, boxes, crates, boats, waterwheels, and wicker furniture. However, the wood often warps badly in curing unless great care is taken, and it is usually too brittle to be used for heavy construction. The tough, pliable shoots of many species are used to make baskets; the bark of other species is used for tanning. Willows are of great value in checking soil erosion. A few species are ornamental shade trees. *See* FOREST AND FORESTRY; TREE.

James F. Ferry

Wind

The motion of air relative to the Earth's surface. The term usually refers to horizontal air motion, as distinguished from vertical motion, and to air motion averaged over a chosen period of 1–3 min. Micrometeorological circulations (air motion over periods of the order of a few seconds) and others small enough in extent to be obscured by this averaging are thereby eliminated. The choice of the 1- to 3-min interval has proven suitable for the study of the hour-to-hour and day-to-day changes in the atmospheric circulation pattern, and the larger-scale aspects of the atmospheric general circulation.

The direct effects of wind near the surface of the Earth are manifested by soil erosion, the character of vegetation, damage to structures, and the production of waves on water surfaces. At higher levels wind directly affects aircraft, missile and rocket operations, and dispersion of industrial pollutants, radioactive products of nuclear explosions, dust, volcanic debris, and other material. Directly or indirectly, wind is responsible for the production and transport of clouds and precipitation and for the transport of cold and warm air masses from one region to another. *See* ATMOSPHERIC GENERAL CIRCULATION; WIND MEASUREMENT.

Cyclonic and anticyclonic circulation. Each is a portion of the pattern of airflow within which the streamlines (which indicate the pattern of wind direction at any instant) are curved so as to indicate rotation of air about some central point of the cyclone or anticyclone. The rotation is considered cyclonic if it is in the same sense as the rotation of the surface of the Earth about the local vertical, and is considered anticyclonic if in the opposite sense. Thus, in a cyclonic circulation, the streamlines indicate counterclockwise (clockwise for anticyclonic) rotation of air about a central point on the Northern Hemisphere or clockwise (counterclockwise for anticyclonic) rotation about a point on the Southern Hemisphere.

When the streamlines close completely about the central point, the pattern is denoted respectively a cyclone or an anticyclone. Since the gradient wind represents a good approximation to the actual wind, the center of a cyclone tends strongly to be a point of minimum atmospheric pressure on a horizontal surface. Thus the terms cyclone, low-pressure area, or low are often used to denote essentially the same phenomenon. In accord with the requirements of the gradient wind relationship, the center of an anticyclone tends to coincide with a point of maximum pressure on a horizontal surface, and the terms anticyclone, high-pressure area, or high are often used interchangeably.

Cyclones and anticyclones are numerous in the lower troposphere at all latitudes. At higher levels the occurrence of cyclones and anticyclones tends to be restricted to subpolar and subtropical latitudes, respectively. In middle latitudes the flow aloft is mainly westerly, but the streamlines exhibit wavelike oscillations connecting adjacent regions of anticyclonic circulation (ridges) and of cyclonic circulation (troughs).

Although the atmosphere is never in a completely undisturbed state, it is customary to refer to cyclonic and anticyclonic circulations specifically as atmospheric disturbances. Cyclones, anticyclones, ridges, and troughs are intimately associated with the production and transport of clouds and precipitation, and hence convey a connotation of disturbed meteorological conditions.

A more rigorous definition of circulation is often employed, in which the circulation C over an arbitrary area bounded by the closed curve S is given by Eq. (1), where the integration is taken completely

$$C = \oint v_t \, dS \tag{1}$$

around the boundary of the area. Here v refers to the wind at a point on the boundary, the subscript t denotes the component of this wind parallel to the boundary, and dS is a line element of the boundary. The component v_t is considered positive or negative according to whether it represents cyclonic or anticyclonic circulation along the boundary S. In this context, the circulation may be positive (cyclonic) or negative (anticyclonic) even when the streamlines within the area are straight, since the distribution of wind speed affects the value of C. *See* ATMOSPHERE; ATMOSPHERIC WAVES, UPPER SYNOPTIC; CLOUD; GEOSTROPHIC WIND; GRADIENT WIND; PRECIPITATION (METEOROLOGY); STORM.

Convergent or divergent patterns. These are said to occur in areas in which the (horizontal) wind flow and distribution of air density is such as to produce a net accumulation or depletion, respectively, of mass of air. Rigorously, the mean horizontal mass divergence D over an arbitrary area A bounded by the closed curve S is given by Eq. (2), where the integration is taken completely around the boundary of the area. Here ρ is the density of air, v refers to the wind at a point on the boundary, and the subscript n denotes the component of this wind perpendicular to the boundary, and dS is an element of the boundary. The component v_n is taken positive when it is directed outward across the boundary and negative when it is directed inward. Convergence is thus synonymous with negative divergence. If spatial variations of density are neglected, the analogous concept of velocity divergence and convergence applies.

$$D = \frac{1}{A} \oint \rho v_n \, dS \tag{2}$$

The horizontal mass divergence or convergence is intimately related to the vertical component of motion. For example, since local temporal rates of change of air density are relatively small, there must be a net vertical export of mass from a volume in which horizontal mass convergence is taking place. Only thus can the total mass of air within the volume remain approximately constant. In particular, if the lower surface of this volume coincides with a level ground surface, upward motion must occur across the upper surface of this volume. Similarly, there must be downward motion immediately above such a region of horizontal mass divergence.

The horizontal mass divergence or convergence is closely related to the circulation. In a convergent wind pattern the circulation of the air tends to become more cyclonic; in a divergent wind pattern the circulation of the air tends to become more anticyclonic.

Regions which lie in the path of an approaching cyclone are characterized by a convergent wind pattern in the lower troposphere and by upward vertical motion throughout most of the troposphere. Since the upward motion tends to produce condensation of water vapor in the rising air current, abundant cloudiness and precipitation typically occur in this region. Conversely, the area in advance of an anticyclone is characterized by a divergent wind pattern in the lower troposphere and by downward vertical motion throughout most of the troposphere. In such a region, clouds and precipitation tend to be scarce or entirely lacking.

A convergent surface wind field is typical of fronts. As the warm and cold currents impinge at the front, the warm air tends to rise over the cold air, producing the typical frontal band of cloudiness and precipitation. *See* FRONT.

Zonal surface winds. Such patterns result from a longitudinal averaging of the surface circulation. This averaging typically reveals a zone of weak variable winds near the Equator (the doldrums) flanked by northeasterly trade winds in the Northern Hemisphere and southeasterly trade winds in the Southern Hemisphere, extending poleward in each instance to about latitude 30°. The doldrum belt, particularly at places and times at which it is so narrow that the trade winds from the two hemispheres impinge upon it quite sharply, is designated the intertropical convergence zone, or ITCZ. The resulting convergent wind field is associated with abundant cloudiness and locally heavy rainfall. A westerly average of zonal surface winds prevails poleward of the trade wind belts and dominates the middle latitudes of

both hemispheres. The westerlies are separated from the trade winds by the subtropical high-pressure belt, which occurs between latitudes 30 and 35° (the horse latitudes), and are bounded on the poleward side in each hemisphere between latitudes 55 and 60° by the subpolar trough of low pressure. Numerous cyclones and anticyclones progress eastward in the zone of prevailing westerlies, producing the abrupt day-to-day changes of wind, temperature, and weather which typify these regions. Poleward of the subpolar low-pressure troughs, polar easterlies are observed.

The position and intensity of the zonal surface wind systems vary systematically from season to season and irregularly from week to week. In general the systems are most intense and are displaced toward the Equator in a given hemisphere during winter. In this season the subtropical easterlies and prevailing westerlies attain mean speeds of about 15 knots (7.7 m/s), while the polar easterlies are somewhat weaker. In summer the systems are displaced toward the pole by 5 to 10° of latitude and weaken to about one-half their winter strength.

When the pattern of wind circulation is averaged with respect to time instead of longitude, striking differences between the Northern and Southern hemispheres are found. On the Southern Hemisphere, variations from longitude to longitude are relatively small due to the predominance of ocean, and the averaged pattern is described quite well in terms of the zonal surface wind belts. On the Northern Hemisphere there are large differences from longitude to longitude because of the ocean-continent contrast. In winter, for example, the subpolar trough is mainly manifested in two prominent low centers, the Icelandic low and the Aleutian low. The subtropical ridge line is drawn northward in effect over the continents and is seen as a powerful and extensive high-pressure area over Asia and as a relatively weak area of high pressure over North America. In summer the Aleutian and Icelandic lows are weak or entirely absent, while extensive areas of low pressure over the southern portions of Asia and western North America interrupt the subtropical high-pressure belt. *See* CLIMATOLOGY; MONSOON METEOROLOGY.

Upper air circulation. Longitudinal averaging indicates a predominance of westerly winds. These westerlies typically increase with elevation and culminate in the average jet stream, which is found in lower middle latitudes near the tropopause at elevations between 35,000 and 40,000 ft (10.7 and 12.2 km). The subtropical ridge line aloft is found equatorward of its surface counterpart and easterlies occur at upper levels over the equatorward portions of the trade wind belts. In high latitudes, weak westerlies aloft are found over the surface polar easterlies. Seasonal and irregular fluctuations of the circulation aloft are similar to those which characterize the surface winds. *See* JET STREAM.

Variation of wind with height. The rate of change of the wind vector with respect to height is called the wind shear. When the wind direction changes in a clockwise sense with increasing height, the wind is said to be veering with height. For example, if a southerly wind at the surface becomes southwesterly a kilometer above the ground, and westerly in the midtroposphere, the wind veers with height. If the geostrophic wind veers with height, then as a consequence of hydrostatic balance, there is warm advection, that is, warmer air is being blown toward colder air. Veering with height often occurs east of surface cyclones, and west of surface anticyclones in midlatitudes. When the wind direction changes in a counterclockwise sense with increasing height, the wind is said to be backing with height. For example, if a northerly wind at the surface becomes northwesterly a kilometer above the ground, and westerly in the midtroposphere, the wind backs with height. If the geostrophic wind backs with height, then there is cold advection, that is, colder air is being blown toward warmer air. Backing with height often occurs east of surface anticyclones and west of surface cyclones.

The winds may also veer with height as a result of surface friction. Typically, at the surface there is a component of wind across isobars from higher to lower pressure, while above the surface the wind tends to blow parallel to the isobars.

The variation of wind speed and direction with height has a significant impact on the type of thunderstorms which can form. Severe thunderstorms and tornadoes often form when the winds veer and increase in strength with height. *See* THUNDERSTORM; TORNADO.

The expressions veering and backing are used also to describe the change of wind direction with respect to time. For example, north of the track of a low in the Northern Hemisphere the winds back with respect to time, while south of the track of a low the winds veer with time.

Minor terrestrial winds. In this category are circulations of relatively small scale, attributable indirectly to the character of the Earth's surface. One example, the land and sea breeze, is a circulation driven by pronounced heating or cooling of a given area in comparison with little heating or cooling in a horizontally adjacent area. During the day, air rises over the strongly heated land and is replaced by a horizontal breeze from the relatively cool sea. At night, air sinks over the cool land and spreads out over the now relatively warm sea.

Another example is formed by the mountain and valley winds. These result from cooling and heating, respectively, of the mountain slopes relative to the horizontally adjacent free air above the valley floor. During the day, air flows up from the valley along the strongly heated mountain slopes, but at night, air flows down the relatively cold mountain slopes toward the valley bottom. A similar type of descending current of cooled air is often observed along the sloping surface of a glacier. This night-time air drainage, under proper topographical circumstances, can lead to the accumulation of a pool of extremely cold air in nearby valley bottoms.

Local winds. These commonly represent modifications by local topography of a circulation of large

scale. They are often capricious and violent in nature and are sometimes characterized by extremely low relative humidity. Examples are the mistral which blows down the Rhone Valley in the south of France, the bora which blows down the gorges leading to the coast of the Adriatic Sea, the foehn winds which blow down the Alpine valleys, the williwaws which are characteristic of the fiords of the Alaskan coast and the Aleutian Islands, and the chinook which is observed on the eastern slopes of the Rocky Mountains. Local names are also given in some instances to currents of somewhat larger scale which are less directly related to topography. Examples of this type of wind are the norther, which represents the rapid flow of cold air from Canada down the plains east of the Rockies and along the east coast of Mexico into Central America; the nor' easter of New England, which is part of the wind circulation about intense cyclones centered offshore along the Middle Atlantic coastal states; and the sirocco, a southerly wind current from the Sahara which is common on the coast of North Africa and sometimes crosses the Mediterranean Sea. *See* CHINOOK; SIROCCO.

Frederick Sanders; Howard Bluestein

Bibliography. J. A. Dutton, *The Ceaseless Wind: An Introduction to the Theory of Atmospheric Motion*, 1976, reprint 1986; J. R. Holton, *An Introduction to Dynamic Meteorology*, 3d ed., 1992; J. M. Wallace and P. V. Hobbs, *Atmospheric Sciences: An Introductory Survey*, 1977.

Wind measurement

The determination of three parameters: the size of an air sample, its speed, and its direction of motion. Air movement or wind is a vector that is specified by speed and direction; meteorological convention indicates wind direction is the direction from which the wind blows (for example, a southeast wind blows toward the northwest). Anemometers measure wind speed, while wind vanes indicate direction. On average, the wind blows horizontally over flat terrain; however, gusts, thermals, cloud outflows, and many other conditions have associated with them significant short-term vertical wind components. While research wind instruments typically measure both horizontal and vertical air movement, operational and personal wind sensors measure only the horizontal component.

There are many types of wind measurement instruments. In-situ devices measure characteristics of air in contact with the instrument; often they are referred to as immersion sensors because they are immersed in the fluid (air) they measure. Remote wind sensors make measurements without physical contact with the portion of the atmosphere measured. Active remote sensors emit electromagnetic (for example, light or radio waves) or sound waves into the atmosphere and measure the amount and nature of the electromagnetic or acoustic power returned from the atmosphere.

Sampling and measurement principles. Various attributes of instrument performance are used to characterize or compare different instruments. Precision, accuracy, and resolution are perhaps the most widely used attributes. Accuracy refers to the degree to which an instrument can measure the true or correct value. Instrument resolution is the smallest change in an atmospheric variable that can be measured—an instrument's resolution is the same as or better than its accuracy. Precision refers to the ability of an instrument to repeatedly measure the same atmospheric state many times. An instrument may be very precise, yet not be particularly accurate. Errors of imprecision are random—they do not occur in a preferred direction away from the true value. Instrument bias refers to measurement errors that consistently underestimate or overestimate the true value. Threshold response is the value of an atmospheric variable at which an instrument begins to respond. A mechanical anemometer may not register a response until the wind speed exceeds a certain value, for example, 0.5 m/s (1 mi/h). At this threshold speed, the force on the anemometer's cups or propeller is able to exceed the friction in the instrument's bearings. Other useful attributes of wind instruments are time constant and distance constant. Time constant is the length of time that the instrument requires to respond to an instantaneous change in atmospheric state; and it is expressed as the time required to indicate 63% of the actual change. Distance constant refers to the amount of air movement (meters) required to indicate a similar amount of change in either wind speed or wind direction.

The quality of atmospheric measurements depends not only on the performance attributes of the instruments but also on the sampling and siting characteristics. The representativeness of a measurement can be affected significantly by the location or siting of the instrument. A low-level wind measurement, for example, can represent the general wind flow pattern of the region, or it can represent the disturbance caused by nearby buildings or vegetation; both measurements are representative, but of very different conditions. Similar considerations exist for all in situ measurements and many remote measurements. Sampling characteristics refer to the manner in which the output signal from the instrument is processed. Sampling rate is the frequency at which the output signal is detected or observed. A wind sock may be sampled visually (that is, observed) at a rate of once per hour, whereas the sampling rate for the digital output of a sonic anemometer may be 20 per second. The length of time that samples are accumulated and subsequently processed is the sampling period. In sampling a phenomenon that has an inherent periodicity, the sampling rate must be greater than the frequency of the atmospheric fluctuations in order to avoid aliasing. If the sampling is aliased, high-frequency phenomena will be indicated as an unknown enhancement of lower-frequency processes. Often the sampled signal is filtered to remove certain frequencies from the processed data set.

High-frequency fluctuations may represent instrument or electronic noise and are removed with a low-pass filter. Low-frequency fluctuations, such as long-term trends, are removed with a high-pass filter. Using a combination of low- and high-pass filters constitutes a band-pass filter that passes the intermediate frequencies.

Ground-based in-situ instruments. There are three principal categories of ground-based in-place wind sensors: mechanical, pressure, and sonic. Mechanical systems are the most common. Virtually all anemometers determine wind speed by measuring the rotational speed of a propeller or a circular array of cups. Wind direction is measured in mechanical devices by the direction of a vertically oriented vane or fin mounted on a horizontal shaft. In a steady wind, the forces on the two sides of the vane are equal. As the wind changes direction, the force on the upwind or windward side of the vane initially exceeds that on the downwind or leeward side until the vane rotates into the wind and the forces again are balanced. Most wind instruments measure wind speed and direction separately, although the propeller vane measures the horizontal wind vector by mounting a propeller at the windward end of a horizontal shaft with a vane on its leeward end. Cups, propellers, and vanes have the common (although usually minor) disadvantage that their inertia causes some overresponse to changes in speed or direction.

The pressure anemometer measures the dynamic pressure exerted by the wind. A Pitot tube—an L-shaped piece of circular tubing open at the windward end—is oriented into the airstream and measures the sum of the static atmospheric pressure and the dynamic pressure of the wind. The dynamic pressure is proportional to the square of the wind speed. A manometer or differential pressure transducer measures the pressure difference between the Pitot tube and a reference. Because the pressure differential responds to the wind speed squared, pressure anemometers are very responsive and accurate at high wind speeds but poorly responsive at speeds less than 1–2 m/s (2–4 mi/h). They frequently are used in wind tunnels and on airplanes; when used for ambient measurements, they must be oriented into the wind by mounting on a vane. *See* PITOT TUBE.

The measurement principle of the sonic anemometer is the dependence of the speed of a sound wave on the wind speed. A sound wave traveling with the wind moves with the sum of the speed of sound and the wind speed, while a wave traveling into the wind has a reduced speed. The three components of the wind vector (vertical, north, and east) are measured by using three pairs of sonic transducers to measure the effective sound speed in three different directions. Sonic anemometers were traditionally used as research instruments, but they have become operational instruments because of their increased reliability and somewhat reduced costs. *See* ANEMOMETER.

Upper-air in-situ measurements. Vertical profiles (usually called soundings) of winds and temperature, humidity, and pressure from the ground surface to pressure altitudes around 25 hectopascals or millibars (25 km or 15 mi) and higher are obtained operationally with a balloon-borne radiosonde. There are nearly 100 radiosonde stations in the United States that launch radiosondes twice daily at 00Z and 12Z (Greenwich mean time or universal standard time) and more than 700 global stations that launch at least one radiosonde daily. The radiosonde consists of expendable sensors, electronics, and radio transmitter to telemeter the measurements back to a receiver at the launch station. Winds are not measured directly, but are determined from the movement of the radiosonde suspended beneath the balloon. One class of wind measurement techniques actually tracks the balloon with one of three methods: (1) an optical system using a theodolite tracks the balloon's azimuth and elevation; (2) a radio theodolite tracks the radio signal from the radiosonde transmitter; and (3) radar systems track a radar retro reflector suspended from the balloon to obtain slant range, azimuth, and elevation. The first two methods require knowledge of the balloon height, which is determined from the pressure measurement. Optical theodolite measurements require clear conditions between the ground station and the balloon. All three methods suffer form a loss of accuracy as the distance to the balloon increases and its elevation angle decreases. *See* SURVEYING INSTRUMENTS; TELEMETERING.

The second class of radiosonde wind measurement techniques uses any of two different radio navigation systems. One system uses the Loran-C radio navigation system. A transponder on the balloon receives the navigation (time-of-arrival) signals and rebroadcasts them back to the ground station, where the wind is determined by tracking the change in position of the balloon. This system requires a pressure measurement to determine the height of the balloon. Another type of navigation-based wind-finding technique uses the Global Positioning System (GPS). It can measure very accurately the actual velocity of the radiosonde, and with some GPS radiosonde receivers it is also possible to measure latitude, longitude, and altitude. Two major advantages of these devices are the high accuracy and precision of the wind measurements and the worldwide coverage of GPS. *See* AIR NAVIGATION; LORAN; SATELLITE NAVIGATION SYSTEMS.

Dropwindsondes are the airborne counterpart to conventional radiosondes and are used primarily in research studies. They are ejected from aircraft and float to the ground on a parachute, measuring winds in the same way as the various standard radio navigation wind-finding methods.

Remote wind sensors. Advances in active remote sensing techniques have led to the transition of radar wind profilers from research devices to operational instruments. Wind profilers transmit a short pulse of radio-frequency electromagnetic energy into the atmosphere, where a small fraction is scattered back from the atmosphere to the profiler's antenna. The intensity of the energy returned from the atmosphere depends on minor irregularities in the index

of refraction—a measure of the temperature, humidity, and pressure characteristics of the air. These irregularities (sometimes referred to as turbulent eddies) are transported with the wind and so act as tracers of air movement. As a result, the frequency of the returned signal is changed (the Doppler shift) according to the speed of the wind along the pointing (radial) direction of the transmitted electromagnetic pulse. Measuring the Doppler shift of the returned signal provides a measure of the radial wind speed component as a function of distance from the antenna (that is, height above the ground). The height increment in which an independent wind speed measurement can be made is equal to one-half the length of the transmitted radio-frequency pulse. By alternately pointing the transmitted beam in various directions, the three-dimensional wind vector can be resolved over successive height intervals (ranges). Some wind profilers can resolve the wind within height intervals about 60 m (200 ft). All wind profilers operate in clear air. The lower-frequency profilers (50 and 400 mhz) are insensitive to rain, while higher-frequency profilers (21 GHz) respond to both clear air and rain; in some cases it is possible to estimate both the wind speed and the rain-drop sizes.

A newer type of radar of radar wind profiler is the spaced antenna system. It measures the wind by sensing the shift in the returned radio signal across a horizontal array of antennas. Spaced antenna systems have the advantage that the wind vector is measured from the return of single, vertically pointed (transmitted) beam, thereby decreasing both the sampling period and the sampled volume of air. Beam-pointing Doppler systems, on the other hand, measure the wind over a broad cone of air formed by three or five different beams. The sampling period required to obtain a representative vertical profile is of order one minute with the spaced antenna profiler, considerably shorter than the Doppler beam-swinging systems.

Lidar is very similar to radar except that it uses a laser to emit pulses of light, which are scattered by small particles in the air. Sodar (sound detection and ranging) refers to wind profilers that use an acoustic pulse to detect the wind in much the same way as radar wind profiles. The acoustic pulses are scattered back to the receiver on the ground by temperature fluctuations in the air. The height coverage with sodar is typically less than with radar or lidar—usually no more than a few kilometers—but the height resolution can be considerably better; mini-sodars can resolve winds over height intervals of a few meters. A major disadvantage of sodar is the irritation of the noise of the acoustic pulses. Unlike the radar wind profiles that are being used as part of operational weather observing systems, sodar and lidar remain primarily research instruments. *See* LIDAR; METEOROLOGICAL INSTRUMENTATION; METEOROLOGICAL RADAR; WIND.

Walter F. Dabberdt

Bibliography. F. V. Brock (ed.), *Meteorological Measurement Systems*, 1995; F. Dobson, L. Hasse, and R. Davis, *Air-Sea Interaction*, 1980; L. J. Fritschen and L. W. Gay, *Environmental Instrumentation*, 1979; D. H. Lenschow (ed.), *Probing the Atmospheric Boundary Layer*, 1986; A. Wexler, *Measurement of Humidity in the Free Atmosphere near the Surface of the Earth: Meteorological Observations and Instrumentation*, Meteorol. Monog. 33, American Meteorological Society, 1970; World Meteorological Organization, *Guide to Meteorological Instruments and Methods of Observation*, WMO 8, 1983.

Wind power

The extraction of kinetic energy from the wind and conversion of it into a useful type of energy: thermal, mechanical, or electrical. Wind power has been used for centuries.

Early development. Windmills were used extensively in the Middle East by the eleventh century and in Europe by the thirteenth century. By the fourteenth century the Dutch used them for draining marshes and lakes. By the end of the nineteenth century, in Denmark there were about 2500 electricity-producing windmills with a generating capacity of about 30 MW, and about 4600 windmills being used for agricultural applications. Farmers and ranchers in the western United States began using small water-pumping windmills in the midnineteenth century. These wind machines were simple durable multibladed turbines with a high starting torque. An estimated 6.5 million machines were sold between 1880 and 1930, and in the 1930s the annual energy output was equivalent to 10^9 kWh per year. Beginning about 1925, small 0.2–3 kW horizontal-axis machines supplying dc electricity to operate appliances, usually at 32 V, were used in rural areas of the United States (**Fig. 1**). The creation of the Rural Electrification

Fig. 1. A small dc horizontal-axis wind machine for rural use.

Administration in 1936 soon brought electrical service to these rural areas and doomed the widespread use of the small machines.

In 1941, a 175-ft-diameter (53-m) 1250-kW-capacity wind turbine was added to the generating system of Central Vermont Public Service Corporation. The machine was located on a 2000-ft (600-m) knob near Rutland, Vermont, and operated until 1945, when one of the blade spars broke and a blade was thrown 750 ft (230 m). The machine was scrapped after the accident, and for all practical purposes the development of wind power in the United States was discontinued until the early 1970s.

Energy in the wind. It has been estimated that the total wind power in the atmosphere averages about 3.6×10^{12} kW, which is an annual energy of about 107,000 quads (1 quad = 2.931×10^{11} kWh). Obviously, only a fraction of this wind energy can be extracted, estimated to be a maximum of 4000 quads per year. The power P in the wind per unit area is given by Eq. (1), where V is wind speed in m/s

$$P = \frac{\rho V^3}{2} \qquad \text{W/m}^2 \tag{1}$$

and ρ is the air density in kg/m^3, which varies from about 1.225 at sea level to 75% this value at 11,000 ft (3000 m) elevation. According to what is commonly known as the Betz limit, a maximum of 59% of this power can be extracted by a wind machine. Practical machines actually extract from 5 to 45% of the available power. Because the available wind power varies with the cube of wind speed, it is very important to find areas with high average wind speeds to locate wind machines. Examples of good areas in the United States are certain mountain passes and ridges in California and Hawaii, and the flat plains in the midwestern regions of the United States. *See* WIND.

Wind-produced electricity. Most research on wind power has been concerned with producing electricity. This effort restarted in the United States in the early 1970s. At present, nearly all the electricity produced in the United States is generated by fossil fuel plants (oil, natural gas, and coal), nuclear plants, and hydroelectric plants. There is concern about fossil fuel generation because the carbon dioxide produced when fossil fuels are burned contributes to the greenhouse effect. Disposal of radioactive waste and concerns about accidents have greatly diminished the expected role of nuclear power, and there is limited hydroelectric power left available for development. The best new sources are conservation and renewable generation. *See* ELECTRIC POWER GENERATION; ENERGY SOURCES; NUCLEAR POWER.

Wind power is a renewable energy source that has virtually no environmental problems. However, wind power has limitations. Wind machines are expensive and can be located only where there is adequate wind. These high-wind areas may not be easily accessible or near existing high-voltage lines for transmitting the wind-generated energy. Another disadvantage occurs because the demand for electricity varies with time, and electricity production must follow the demand cycle. Since wind power varies randomly, it may not be available when needed. The storage of electrical energy is difficult and expensive, so that wind power must be used in parallel with some other type of generator or with nonelectrical storage. Wind power teamed with hydroelectric generators is attractive because the water can be used for energy storage, and operation with underground compressed-air storage is another option. *See* ENERGY STORAGE; PUMPED STORAGE; WATERPOWER.

At present, the best locations for wind power are sites where average wind speeds are high and where nearby conventional generation has a high incremental cost, so that when wind power is available it can produce large cost savings. This combination of factors occurs in California. As a result, by the end of 1985 there were about 13,370 wind generators operating in that state with a total capacity of approximately 1120 MW (a large nuclear plant is about 1100 MW). Most of these machines are grouped into wind farms in mountain passes.

Types of wind machines. The most common type of wind turbine for producing electricity has a horizontal axis, with two or more aerodynamic blades mounted on the horizontal shaft. The wind machine shown in Fig. 1 is a small horizontal-axis turbine. With a horizontal-axis machine, the blade tips can travel at several times the wind speed, which results in a high efficiency. The blade shape is designed by using the same aerodynamic theory as for aircraft. *See* PROPELLER (AIRCRAFT).

Several large wind machines were built by the National Aeronautics and Space Administration under the sponsorship of the Department of Energy. The Mod-0 had a two-bladed 125-ft-diameter (38-m) rotor, which rotated at 40 revolutions/min and drove a 100-kW synchronous generator. The rotor height was 100 ft (30.5 m). The Mod-1 had a two-bladed, 202-ft-diameter (61.5-m), 35-revolutions/min rotor, which drove a 2000-kW synchronous generator. The hub height was 140 ft (42.7 m). The Mod-2 (**Fig. 2**) had a two-bladed, 300-ft (91.4 m), 17.5-revolutions/min rotor driving a 2500-kW synchronous generator. The hub height was 200 ft (61 m). The Mod-0 and a later version named the Mod-0A operated satisfactorily. The Mod-1 and Mod-2 had serious problems.

Many other small horizontal-axis machines are produced. The wind farms in California are mainly composed of these machines, with sizes ranging from a few kilowatts to over 200 kW per machine. Vertical-axis machines have also been investigated. The most common vertical-axis machine is the Darrieus or egg-beater. Two or more curved blades are attached to each end of a vertical shaft. The Darrieus machine has not been as widely used as the horizontal-axis machine for producing electricity.

Estimating energy output. The energy output from a wind machine can be estimated from the graph of machine power output versus wind speed and the wind characteristics at a site. A simplified graph of

Fig. 2. The 2500-kW Mod-2 wind machine at Goodnoe Hills near Goldendale, Washington.

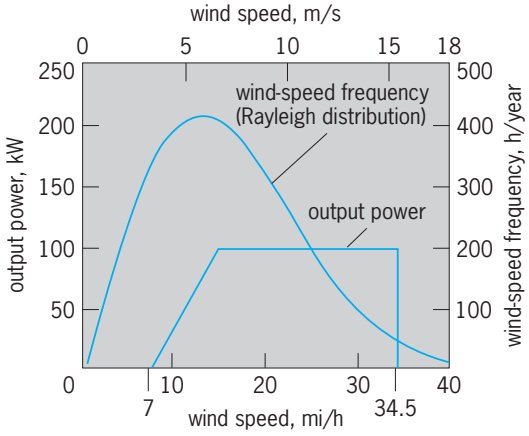


Fig. 3. Approximate Mod-0 power output curve and Rayleigh wind-speed distribution for average wind speed of 16 mi/h (7.2 m/s). The sharp cutoff in output power at a wind speed of 34.5 mi/h (15.4 m/s) is caused by the wind machine shutting down for safety at high speeds.

output power as a function of wind speed measured at a height of 30 ft (9.1 m) for the Mod-0 wind machine is shown in **Fig. 3**.

Often the wind characteristic at a site is simply specified by the average wind speed. However, wind speed V is a random variable about this average value. The Rayleigh distribution, given in Eq. (2), often

$$f(V) = \frac{2V}{c^2} e^{-(V/c)^2} \tag{2}$$

approximates the wind-speed variation. Here, $f(V)$ is the frequency with which various wind speeds occur, and c is related to the average wind speed V_{av} by Eq. (3). The Rayleigh distribution of wind-speed

$$c = \frac{2V_{av}}{\sqrt{\pi}} \tag{3}$$

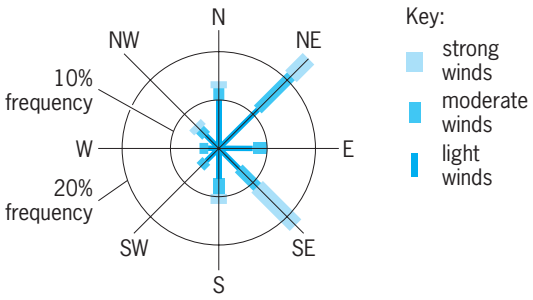
frequencies for $V_{av} = 16$ mi/h (7.2 m/s) is shown in Fig. 3. This curve gives the hours per year the wind is in each 1-mi/h range. For example, it can be seen that the wind would be at 20 mi/h about 320 h per year. *See* PROBABILITY; STATISTICS.

The yearly wind-machine output can be determined by multiplying the output at each wind speed by the number of hours that the wind is at that speed and then summing for all wind speeds. For the example of Fig. 3, it is calculated that this wind machine would produce 590,000 kWh per year if it was available for operation 100% of the time. This means the wind machine has an average output of 67% of its maximum, which is a very high percentage. However, a wind site with an average wind speed of 16 mi/h (7.2 m/s) at 30 ft (9.1 m) elevation is also an exceptionally good wind site. Gary Thomann

Bibliography. T. Ackerman (ed.), *Wind Power in Power Systems*, 2005; T. Burton et al., *Wind Energy Handbook*, 2001; J. F. Manwell et al., *Wind Energy Explained*, 2002; D. A. Spera (ed.), *Wind Turbine Technology: Fundamental Concepts of Wind Turbine Engineering*, 1994.

Wind rose

A diagram in which statistical information concerning the direction and speed of the wind at a particular location may be conveniently summarized. In the standard wind rose a line segment is drawn in each of perhaps eight compass directions from a common origin (see **illus.**). The length of a particular



Standard wind rose.

segment is proportional to the frequency with which winds blow from that direction. Parts of a given segment are given various thicknesses, indicating frequencies of occurrence of various classes of wind speed from the given direction. *See* WIND MEASUREMENT. Frederick Sanders

Wind stress

The drag or tangential force per unit area exerted on the surface of the Earth by the adjacent layer of moving air. Erosion of ground surfaces and the production of waves on water surfaces are manifestations of wind stress. Surface wind stress determines the exchange of momentum between the Earth and the atmosphere and exerts a strong influence on the typical variation of wind through the lowest kilometer of the atmosphere. Estimated values of the surface wind stress range up to several dynes per square centimeter (0.1 pascal), depending on the nature of the surface and the character of the adjacent airflow. *See* METEOROLOGY.

Internal horizontal stresses. Significant stresses arise within the lower atmosphere because of the strong shear of the wind between the slowly moving air near the ground and the more rapidly moving air a kilometer above and because of the turbulent nature of the airflow in this region. The turbulent eddies referred to here have characteristic dimensions ranging up to a few hundreds of meters.

The effectiveness of this turbulent viscosity in the transferring of momentum from level to level may, in favorable circumstances, be a million times as great as the effectiveness of purely molecular viscosity. These turbulent viscous stresses exert an indirect effect on the surface wind stress by effecting the transfer of large amounts of momentum from higher levels down to levels adjacent to the surface of the Earth.

The torque exerted by the surface wind stress averaged over the entire Earth and over a sufficiently long period of time must be equal to zero. Otherwise the net torque acting between the Earth and the atmosphere would tend to alter the rate of rotation of the Earth about its axis and to alter the mean wind circulation of the atmosphere. In the atmosphere the angular momentum supplied, in effect, by the Earth in regions of easterly surface flow balances the angular momentum drained by the Earth in regions of westerly wind flow, so that the average torque exerted by the surface wind stress is indeed zero, or very nearly so. The average magnitude of the surface wind stress is, of course, not equal to zero. It determines the average rate at which kinetic energy of the winds is dissipated by surface friction. *See* ANGULAR MOMENTUM; ATMOSPHERE.

Wind pressure. This is the force exerted by the wind per unit area of solid surface exposed normal to the wind direction and is also known as dynamic pressure. In contrast to shearing stresses, the wind pressure arises from the difference in pressure between the windward and lee sides of the exposed surface. Wind pressure thus represents a substantial force when the wind speed is high. *See* WIND.

Frederick Sanders

Stress over sea. The drag or tangential force of the wind on the sea is expressed in units of dynes per square centimeter or micronewtons per square meter but is normally taken to represent the mean drag over an undefined area, perhaps several kilometers square, containing many waves. It is usually related to an appropriate time and space average of the wind near the sea surface (at 10 m above the mean level, for example).

Over most surfaces overland, in the absence of vertical density gradients, the drag varies as the square of the wind speed and is expressed, following aerodynamic usage, by the equation shown below, where

$$\tau = C_D \rho v^2$$

τ is the shearing stress, ρ is the air density, and v is the wind speed at a specified height, say 10 m. For a fixed (uniform) surface geometry, or roughness, C_D is a constant known as the drag coefficient. *See* AERODYNAMIC FORCE.

Over a water surface a drag coefficient is defined in the same way, but since the surface geometry varies with the wind speed, C_D is to be regarded as a variable. The drag coefficient over the sea is an important quantity in both meteorology and oceanography since it relates the wind speed to the drag, which generates ocean waves, drives the ocean currents, and sets the scale of the atmospheric turbulence that transfers water vapor and heat from the ocean to the atmosphere to provide the energy for clouds and weather systems. *See* ATMOSPHERIC GENERAL CIRCULATION; MARITIME METEOROLOGY; OCEAN CIRCULATION; OCEAN WAVES.

The drag coefficient of the sea surface depends on the wave field and on the turbulent structure of the flow in the air and the water. Present knowledge of the complicated fluid mechanics involved is not sufficient to allow theoretical calculation of it.

Methods of measurement. Many methods have been used to measure the drag of the wind on the sea, but they are of four main types. The first technique attempts to get information from the wind tilt of enclosed bodies of water. When wind blows over a lake, the surface becomes tilted upward toward the lee end. Measurements of the mean slope, which typically would be of the order of 1 cm in 10 km, can provide estimates of the wind drag. There are theoretical difficulties associated with end effects, but the major difficulties are practical: steady isothermal conditions are required, and although the surface waves can be filtered out to permit observations of the necessary accuracy, it is often hard to eliminate the longperiod oscillations of the whole body of water, which are known as seiches. *See* SEICHE.

The second method involves the measurement of the vertical of the mean wind in the lowest few meters of air over the sea. When air and water are at the same temperature, wind speed increases as the logarithm of the height, and its rate of increase provides a measure of the wind drag. This method is empirical, but it is well established; the difficulties are of engineering rather than of science. It is difficult to get instruments of the necessary precision (the wind speed has to be measured to about 1 cm/s) and even harder to expose them over water in such a way that their supports do not interfere with the wind or the waves or the currents. For this reason many

of the observations have been made in lakes and in shallow water, although some workers have developed techniques for making wind measurements accurately enough from small ships or buoys.

The third technique again involves measuring the vertical distribution of the mean wind, but in this case to greater heights. In steady conditions the drag of the wind on the sea can be inferred from the component of the wind across the isobars; therefore, for this method accurate measurements of the distribution of atmospheric pressure, as well as accurate wind observations, are necessary. This method does not work well in middle latitudes because in those areas the wind blows at only a small angle to the isobars in the lowest kilometer or so. It works better in the trade winds, where attempts to use it have met with moderate success.

The fourth technique again requires detailed wind measurements, in this case detailed in time by using an instrument of rapid response, reacting, for example, in about one-tenth of a second. Measurements are made of both the horizontal and vertical components of the wind, and their mean product provides the drag. There are again some theoretical difficulties in the interpretation of the results, but they can probably be overcome. The major difficulties are again of getting suitable instruments suitably exposed, and of the large amount of numerical analysis that has to be done to get the results. On the other hand, the method provides detailed information on the structure of the turbulent flow, and it is being increasingly used. A further attraction of this method is that estimates of evaporation and direct heat transfer can be made by measuring fluctuations of humidity and temperature as well as those of the vertical velocity component. The drag can also be inferred from observations of the fluctuations at high frequencies by deducing the energy dissipation, with the assumption that the smallscale turbulence is isotropic.

Measured values. There is substantial agreement that the drag of the wind on the sea is small relative to that of a fixed soil surface with the same geometry. It is largely independent of the fetch and so seems to depends less on the larger waves than on the short waves and ripples. Surface-active agents, which affect the shortest waves, may therefore be important.

J. R. Garratt provided evidence that C_D increases slowly with wind speed υ from about 10^{-3} at 5 m · s^{-1} to about 2×10^{-3} at 20 m · s^{-1}, possibly reaching 4×10^{-3} at 50 m · s^{-1}.

This is consistent with the notion that the effective aerodynamic roughness of the sea surface is itself determined by the local shearing stress and the acceleration of gravity. But the mechanism remains obscure. Henry Charnock

Bibliography. S. L. Hess, *Introduction to Theoretical Meteorology*, 1959, reprint 1979; D. Houghton and F. Sanders, *Weather at Sea*, 3d ed., 1998; F. K. Lutgens and E. J. Tarbuck, *The Atmosphere: An Introduction to Meteorology*, 8th ed., 2000; A. Miller and J. C. Thompson, *Elements of Meteorology*, 6th ed., 1999; P. Sachs, *Wind Forces in Engineering* 2d ed., 1978; S. D. Smith, Coefficients for sea surface wind stress, heat flux and wind profiles as a function of wind speed and temperature, *J. Geophys. Res.*, 93(C12):15467–15472, 1988.

Wind tunnel

A duct in which the effects of airflow past objects can be determined. The steady-state forces on a body held still in moving air are the same as those when the body moves through still air, given the same body shape, speed, and air properties. Scaling laws permit the use of models rather than full-scale aircraft. Models are less costly and may be modified more easily than aircraft, and conditions may be simulated that would be impossible or dangerous in flight.

Related aerodynamic research equipment includes ballistic guns, drop models, rockets, whirling arms, rocket-powered sleds, and flight tests. The wide range of problems studied requires a similarly wide range of specialized wind tunnels. *See* AIRCRAFT TESTING.

Uses and Methods

Determinations commonly made with wind tunnels are as follows:

- Drag of airplanes and missiles
- Lifting characteristics of winged vehicles or lifting bodies
- Static stability of aircraft, including missiles
- Dynamic stability derivatives of aircraft
- Torques required to deflect control surfaces
- Pressure distributions to determine air loads
- Flutter characteristics of flexible aircraft
- Distribution and rate of heat transfer to aircraft parts for cooling and structural design
- Performance of air-breathing engines and inlet airflow
- Performance of propellers
- Conditions for safe release of bombs and missiles
- Effect of wind on buildings, bridges, signs, automobiles, and other nonflying structures
- Nature of smoke flow from factories or ships
- Nature of wind flow around geographic features

Test conditions. The tunnels in which experiments are made should closely match the conditions of full-scale flight, but it is not always possible to match all the scale parameters. An attempt is always made to match scale and the effects of compressibility of the air. *See* DIMENSIONAL ANALYSIS; DIMENSIONLESS GROUPS; DYNAMIC SIMILARITY; MODEL THEORY.

If compressibility effects are neglected, scale is duplicated and flow patterns are similar when model tests and full-scale flight are at the same Reynolds number (Re). The Reynolds number is proportional to the product of speed and a characteristic length. *See* REYNOLDS NUMBER.

Compressibility effects are similar when the ratio of remote airspeed to the velocity of sound, called

Mach number (M), is duplicated. *See* MACH NUMBER.

At flight speeds below a Mach number of approximately 0.7, one need be concerned only with the Reynolds number, since the effects of the Mach number are small and may be readily calculated. It is usually sufficient to require that the Reynolds number be greater than 1,500,000. If the flight speed is above M = 0.7, the effects of compressibility are less predictable and the Mach number must be matched. If the Reynolds number exceeds 4,000,000 the scale match is usually adequate. However, research has identified problems in the transonic speed range associated with a Reynolds number above 100,000,000. No wind tunnels are available that can match that combination of Reynolds number and M. *See* TRANSONIC FLIGHT.

Special tests require the matching of other similarity parameters. Where dynamic characteristics are desired, the mass and inertia must be scaled with the forces to produce similar flight paths. If the model to be simulated is flexible, its elastic properties must be matched.

Data taken from models in wind tunnels suffer from extraneous effects due to the structure that supports the model and from the tunnel walls. The effects of support interference are minimized by design and are evaluated by careful experiments. The interference of the walls is calculated from theoretical considerations and is evaluated for each tunnel and type of model.

Measurements. Most data are acquired from wind tunnels through measurement of forces and moments, surface pressures, changes produced in the airstream by the model, local temperatures, and motions of dynamically scaled models, and by visual studies.

Force and moment measurements. A balance system separates and measures the six components of the total force. The three forces, taken parallel and perpendicular to the flight path, are drag, lift, and side force. The three moments about these axes are yawing moment, rolling moment, and pitching moment, respectively.

Surface pressure. Surface pressures are measured by connecting orifices flush with the model surface to pressure-measuring devices. Local air load, total surface load, moment about a control surface hinge line, boundary-layer characteristics, and local Mach number may be obtained from pressure data.

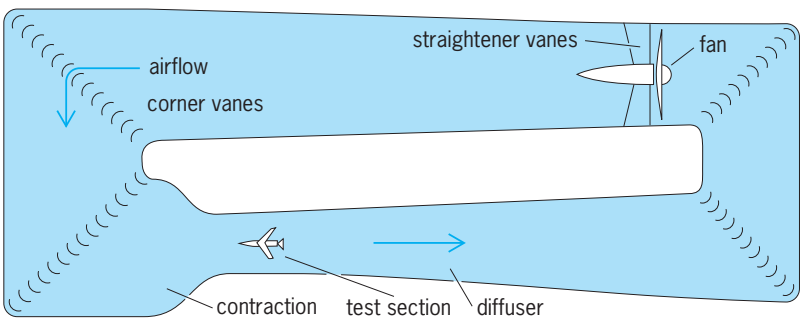


Fig. 1. Outline plan view of low-speed return-flow wind tunnel.

Reaction of model on airstream. Measurements of stream changes produced by the model may be interpreted in terms of forces and moments on the model. In two-dimensional tunnels, where the model spans the tunnel, it is possible to determine the lift and center of pressure by measuring the pressure changes on the floor and ceiling of the tunnel. The parasite drag of a wing section may be determined by measuring the total pressure of the air which has passed over the model and calculating its loss of momentum.

Surface temperature. Measurements of surface temperatures indicate the rate of heat transfer or define the amount of cooling that may be necessary.

Dynamic measurements. In elastically and dynamically scaled models used for flutter testing, measurements of amplitude and frequency of motion are made by using accelerometers and strain gages in the structure. In free-flight models, such as bomb or missile drop tests, data are frequently obtained photographically. *See* ACCELEROMETER; FLUTTER (AERONAUTICS); STRAIN GAGE.

Flow visualization. At low speeds, smoke and tufts are often used to show flow direction. A mixture of lampblack and kerosine painted on the model shows the surface streamlines. A suspension of talcum powder and a detergent in water does the same, is cleaner, and has a more pleasant odor.

At velocities near or above the speed of sound, some flow features may be made visible by optical devices. *See* INTERFEROMETRY; SCHLIEREN PHOTOGRAPHY.

Types

There are various types of wind tunnels, including low-speed tunnels; V/STOL tunnels; transonic, supersonic, and hypersonic tunnels; nonaeronautical tunnels; and hypervelocity tunnels.

Low-speed wind tunnel. This tunnel has a speed up to 300 mi/h (480 km/h). A low-speed tunnel has the essential features of most wind tunnels. These include a device to drive the air, a duct that provides smooth, steady, parallel flow of air in the test section, and instrumentation to measure the desired information. The tunnel may be open-circuit or return-flow (**Fig. 1**). The test section may be closed (with walls), or may be an open jet. When air-burning engines are to be tested, an open-circuit tunnel is appropriate. Most tunnels use a return duct, primarily to ensure smooth flow. The shape of the tunnel, continuously expanding from test section through the fan and around to the settling chamber, is chosen to reduce friction losses in the return circuit and to provide the powerful smoothing effect of a large contracting section just upstream of the test section. Except for special-purpose tunnels, most low-speed tunnels have a closed test section. The propulsion device (fan) causes an increase of pressure to overcome the friction losses in the tunnel circuit. Low-speed tunnels of convenient working-size test section (8×12 ft or 2.5×3.6 m) can test models at adequate Reynolds numbers with speeds up to 250 mi/h (400 km/h).

In a well-designed low-speed tunnel having a closed return, the kinetic energy of the air flowing

through the test section per second may be 6 to 10 times the rate at which energy is being supplied by the fan. This energy ratio (ER) may be used to estimate the power required to drive a tunnel by using the formula given below, where A_{ts} is test-section

$$\text{Kilowatts required} = \frac{13.13}{(ER)} A_{ts} \left(\frac{V}{100} \right)^3 \sigma$$

cross-section area in square meters, V is air velocity in kilometers per hour, and σ is the ratio of the air density to the density at sea level. The energy ratio may be as low as 1 or 2 for an open-circuit tunnel and is typically about 6 to 8 for most low-speed tunnels.

Lifting models experience interference in the tunnel because the walls restrict the downward flow behind the wing. The measured lift would only be achieved in free air at a larger angle of attack, and so a correction (typically less than 10%) is added to the angle of attack. The measured drag is reduced because of the angle-of-attack interference, and a correction which is a product of the lift and the angle correction is added to the measured drag. When the model has a tail, it experiences a different angle interference from that of the wing, and so a correction to the pitching moment is necessary. The drag and pitching moment corrections are of significant magnitude, but are well understood and are used confidently.

Since the model scale is usually smaller than the full-scale airplane, the Reynolds number is not matched and some scale effects appear in the data. Typically, the maximum lift of a wing is too small and the minimum drag is too large. The error can be substantial and is not subject to simple analysis. A large body of experience with these effects permits the engineer to estimate the effects of Reynolds number mismatch.

V/STOL wind tunnel. This type is a newer development of low-speed wind tunnels having a large very-low-speed section to permit testing of aircraft designed for vertical or short takeoff and landing (V/STOL) while operating in the region between vertical flight and cruising flight. In very slow flight, the flow is characterized by large downwash angles from the lifting system. The lift is usually developed by the use of power applied to a rotor, fan, jet, or jet flap. Because the rotating parts or the jet flows operate at high speeds, the flight Mach number becomes the important parameter to match. This requires that tunnel airspeeds equal flight speeds (from 20 to 100 mi/h or 32 to 160 km/h). Large test sections are required to allow the power-driven downwash to leave the model without too much distortion due to the proximity of the tunnel walls.

For economy reasons, most V/STOL tunnels are built with a standard low-speed test section (up to 300 mi/h or 480 km/h) in tandem downstream of the V/STOL test section (up to 100 mi/h or 160 km/h). This two-test-section wind tunnel provides a wide range of test conditions and has advantages in speed control at very low speeds (**Fig. 2**).

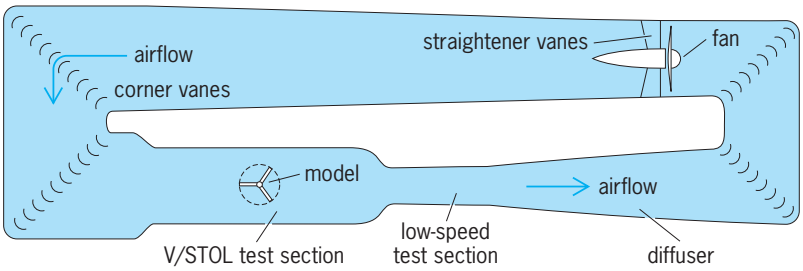


Fig. 2. Combined V/STOL and low-speed wind tunnel with two test sections.

Instrumentation and test methods are similar to those of the low-speed tunnel, but models are usually more complicated because of the power systems required. Wall interference problems are yielding to research, but corrections are still difficult. Research has continued in this area. *See* SHORT TAKEOFF AND LANDING (STOL); VERTICAL TAKEOFF AND LANDING (VTOL).

Transonic tunnel. This type is a high-speed tunnel, capable of testing at speeds near the speed of sound, at Mach numbers from 0.7 to 1.4. In compressible fluid flow in a closed duct, sonic speed can occur only where the cross section is a minimum. Therefore, sonic speed occurs where the model is located. Additional power only causes areas of supersonic flow downstream. *See* COMPRESSIBLE FLOW.

To achieve transonic flow, the test section is surrounded by a large plenum chamber and the test-section walls are vented to this chamber by means of perforations or slots parallel to the airstream. The function of the vents is as follows. Shock waves are formed in the flow near the model when sonic speed is reached. Shock waves are reflected from a solid boundary but with a change of sign (that is, expansion waves instead of pressure waves) from a free boundary. By providing sufficient free surfaces in the form of perforations or slots, reflections which would alter the flow around the model are greatly reduced or canceled. Ventilation has proved to be useful at Mach numbers from 0.9 to 1.2, where reflections are particularly troublesome. The raw data from ventilated test sections are usually used without correction except for scale effect. *See* SHOCK WAVE.

The energy ratio for a transonic tunnel may be as large as 10 at subsonic speeds, and drop to 3 at transonic speed, leading to a power requirement of up to 5000 kW/m^2 of test section.

Supersonic wind tunnel. This tunnel is capable of test speeds corresponding to Mach numbers from 1.0 to 5. For supersonic speeds, the test section must be designed for the particular Mach number desired. When air accelerates past the speed of sound, it expands more rapidly than it accelerates, so that the tunnel must be larger downstream than it is at the minimum section, where the speed is sonic. The final Mach number is uniquely determined by the ratio of the final area to the throat area (**Fig. 3**)

Continuous-flow supersonic tunnels may require 20,000–50,000 kW/m^2 of test section. Most industrial tunnels operate intermittently from energy stored in high-pressure air tanks. The air is discharged

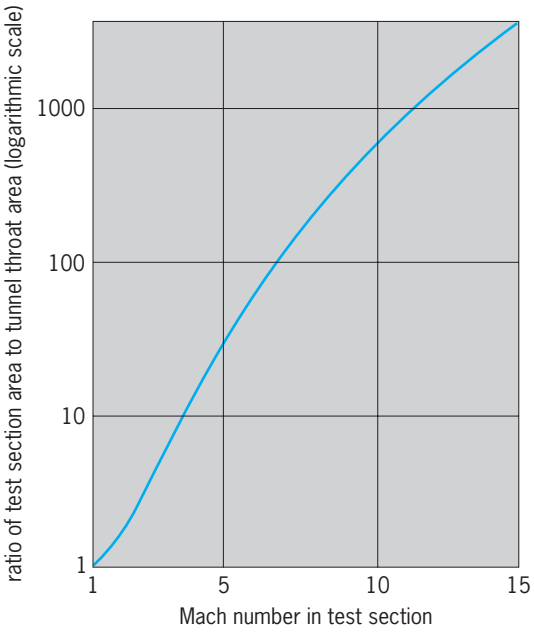


Fig. 3. Ratio of test section areas to throat area for supersonic velocity in a wind tunnel.

through a fast-acting regulator valve to the tunnel and exhausted to the atmosphere (**Fig. 4***a*).

The successful development of rapid recording instrumentation has made short run times possible.

The indraft, or vacuum-driven, tunnel allows atmospheric-pressure air to flow through the wind tunnel into a vacuum tank (Fig. 4*b*). Advantages of the indraft system include a constant (atmospheric) stagnation pressure and temperature, greater safety, and less noise and cost. Advantages of the blowdown type include higher Reynolds numbers and the ability to vary the Reynolds number over a wide range by varying the stagnation pressure.

The principal problems in the design of a supersonic tunnel are to provide sufficient pressure ratio to start and sustain the flow, and to supply adequately dry air. Drying the air to a dew point of about −40°F (−40°C) at atmospheric pressure may be required to avoid condensation shocks and uneven flow.

No wall interference effects are felt in supersonic flow, provided reflecting shock waves do not strike the model. Adjustment of results for scale effects (Reynolds number) must usually be made. *See* SUPERSONIC FLIGHT.

Hypersonic wind tunnel. This type is a supersonic tunnel operating in the range of Mach numbers from 5 to 15. When the Mach number is above 4 or 5, extremely high pressure ratios are required and the temperature drop in the test section is so large that liquefaction of the air would result unless the air were first heated (**Fig. 5**). Most hypersonic tunnels are intermittent types and use both high-pressure stagnation air and a vacuum-discharge tank to obtain the overall pressure ratio required (**Fig. 6**).

Heat is added by using electrical or gas-fired heaters or by passing the air through a bed of ceramic pebbles which have been previously heated. The nozzle may require cooling. The test section usually operates at a very low density, which means low Reynolds number and thick boundary layers.

Data from hypersonic wind tunnels are used directly but are subject to serious Reynolds number effects because of the low density of the flow. *See* HYPERSONIC FLIGHT.

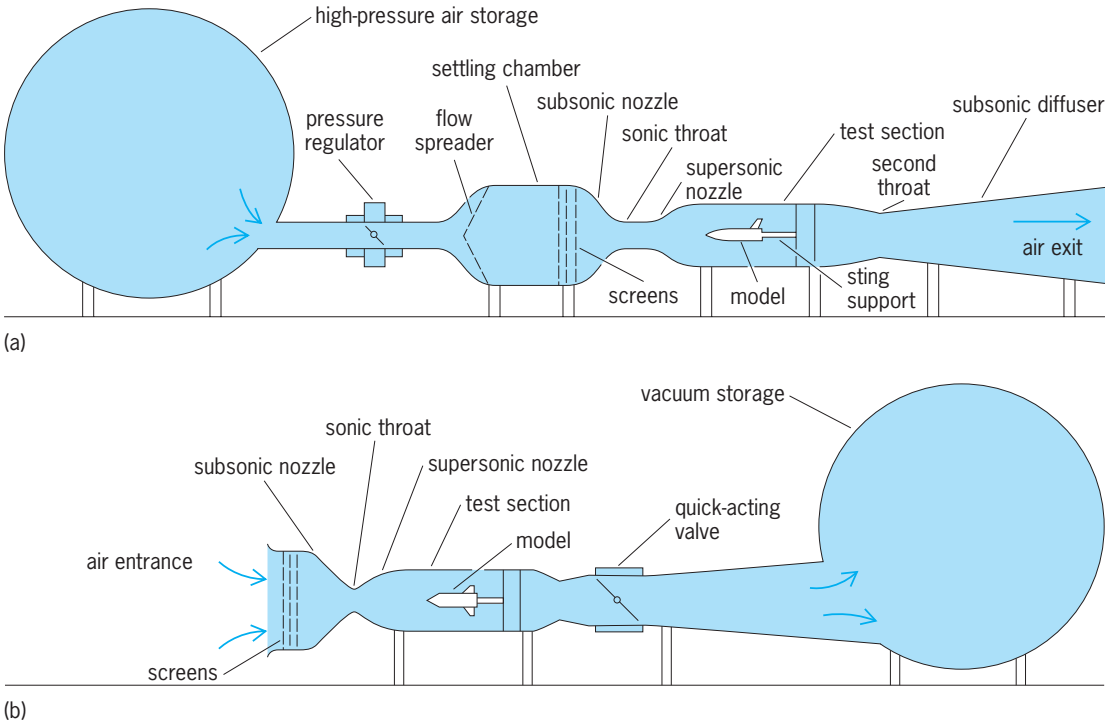


Fig. 4. Elements of (*a*) intermittent supersonic blowdown wind tunnel and (*b*) intermittent indraft supersonic wind tunnel.

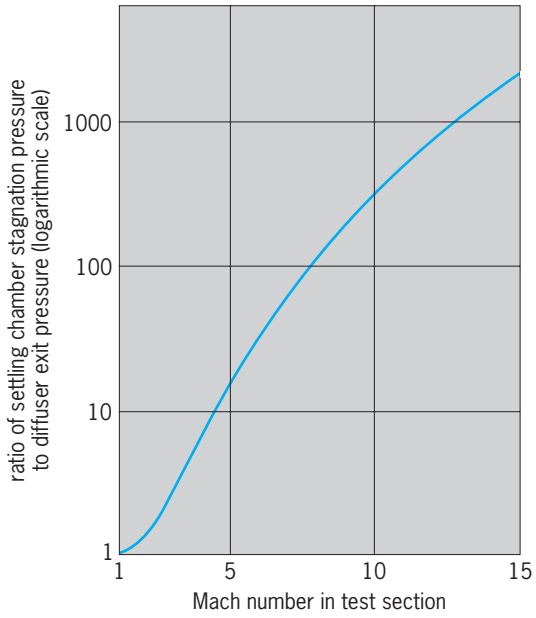


Fig. 5. Pressure ratio required to start and sustain supersonic airflow. These theoretical values assume normal shock wave at the end of the test section.

Nonaeronautical wind tunnels. Testing for the effects of wind on buildings, automobiles, bridges and other nonflying objects often requires wind tunnels of special sizes and shapes.

Meteorological wind tunnels have test sections whose length is 10 to 15 times as long as they are high (**Fig. 7**). In order to develop a thick boundary layer to match the natural wind, roughness elements such as cubes or tall spirelike structures may be located on the tunnel floor upstream of the model. The roof may be adjustable to permit matching of natural zero pressure gradient. These tunnels may be used to measure the forces on buildings, and how the flow around them affects the immediate neighborhood. Dispersal of smoke and other pollutants from factories may be studied. Snowdrift development, wind power generation, and dynamic testing of elastically scaled buildings or other structures are also done in meteorological wind tunnels.

Automobile wind tunnels are often very similar to aeronautical tunnels when force data are needed. For cooling tests, the tunnel may provide a jet of air not much larger than the car itself.

Special tunnels have been built for testing the dynamic character of suspension bridge models in which the test section was 50 times as wide as it was high.

Use of computers. Use of computers is rapidly changing the way in which wind tunnels are operated. The calculations involved in the conversion of raw data, in the form of forces and pressures, to final coefficient form are relatively simple but voluminous. More of this work, called data reduction, has been done by computers as their speed and memory capacity has increased. Advances in computer capability have led to changes from batch processing at the end of each day, to processing at the end of each run, to on-line data reduction on a point-by-point basis. Coefficients are plotted on graphic terminals while a test is in process. The test engineer can observe the progress of the test and can recall previous tests from storage for comparison. *See* COMPUTER GRAPHICS.

The computer is also used to operate the tunnel, automatically following a preprogrammed plan of test. The computer can control the tunnel speed, cause the required changes of pitch and yaw angles, and turn on the data system when specified test conditions are established.

The computer has made possible more detailed analysis of flow in wind tunnels, and has made adaptive wall tunnels possible. When the flow characteristics of a given model in free air can be calculated at a distance corresponding to the tunnel walls, the walls may be adjusted to match the free air streamlines. When this is done, the interference of the tunnel walls vanishes and no correction to the data is needed. The adjustment of the walls may be actual physical motion or virtual movement produced by injecting or extracting air through slots or holes in the walls. The techniques may be used for V/STOL or for transonic testing where corrections may be large and uncertain. Such tunnels may be called smart wind tunnels or adaptive tunnels. *See* DIGITAL COMPUTER.

Robert G. Joppa

Hypervelocity wind tunnels. Hypervelocity wind tunnels are tunnels possessing speed and temperature capabilities greater than those of hypersonic tunnels. They encompass the velocity range greater than 5000 ft/s (1500 m/s at Mach numbers 1 up to about 25) and are necessary for conducting aerodynamic research on vehicles such as ballistic-missile and satellite reentry bodies. *See* ATMOSPHERIC ENTRY.

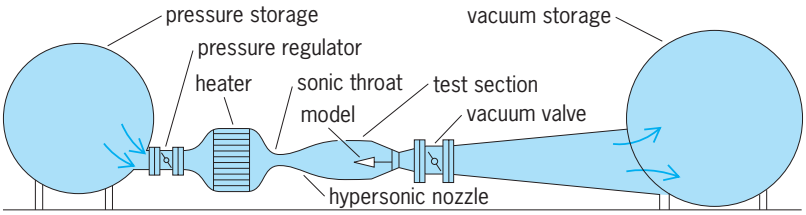


Fig. 6. Hypersonic wind tunnel with both pressure blowdown and vacuum indraft.

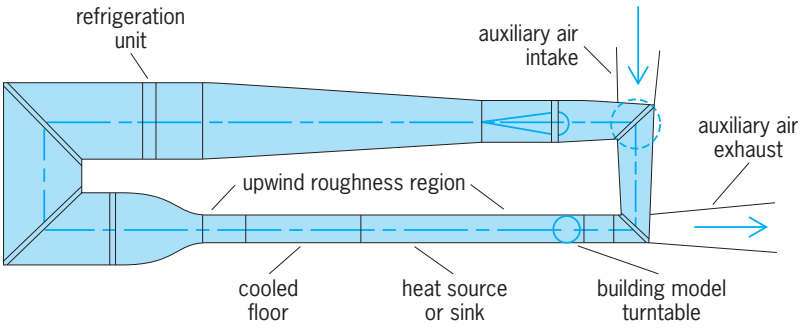


Fig. 7. Meteorological wind tunnel. (*After W. H. Rae, Jr., and A. Pope, Low-Speed Wind Tunnel Testing, 2d ed., Wiley-Interscience, 1984*)

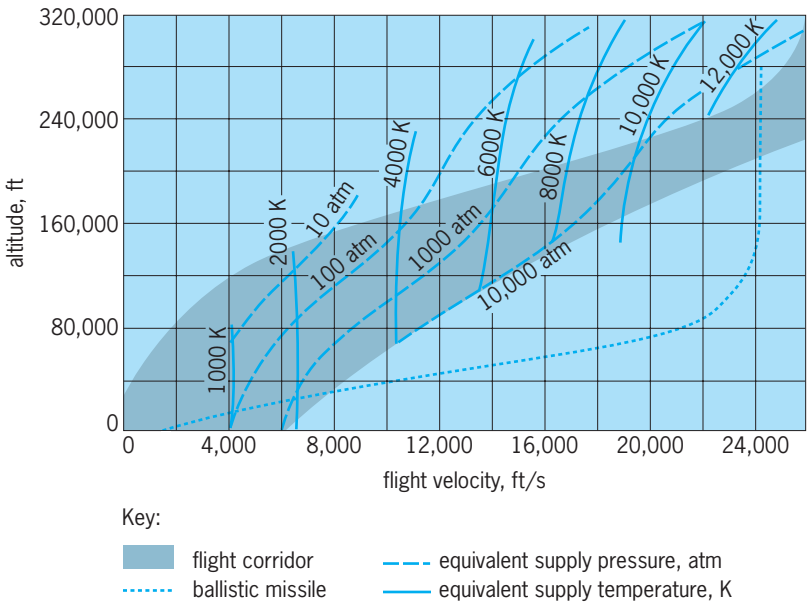


Fig. 8. Wind tunnel simulation of hypervelocity flight requires extremely high supply (stagnation) temperatures and pressures. 1 ft. = 0.3 m; 1 ft/s = 0.3 m/s; 1 atm = 10^2 kPa = 14.7 lb/in.^{2}; $^\circ$F = (K × 1.8) − 460.

Wind-tunnel simulation of hypervelocity flight conditions requires extremely high supply (stagnation) temperatures and pressures, as may be seen in **Fig. 8**. These high supply temperatures are required for expanding the tunnel airflow through the nozzle to hypervelocities in the test section; similarly, high supply pressures are necessary to duplicate flight-altitude ambient pressure in the test section. Hypervelocity tunnels operate at considerably higher supply temperatures and pressures than do hypersonic tunnels.

The "flight corridor" indicated in Fig. 8 represents an approximate altitude-velocity band wherein flight of vehicles supported by aerodynamic lift and centrifugal force can be sustained. The upper boundary of the flight corridor is fixed by the minimum allowable atmospheric density; the lower boundary is fixed by the maximum allowable aerodynamic heating rate. This corridor encompasses the flight spectrum of the proposed hypersonic transport or boost-glide vehicles. A typical trajectory of a ballistic missile from its entry into the Earth's atmosphere is shown in Fig. 8. The missile is able to tolerate the higher heating rates encountered below the corridor of continuous flight because of its short flight duration.

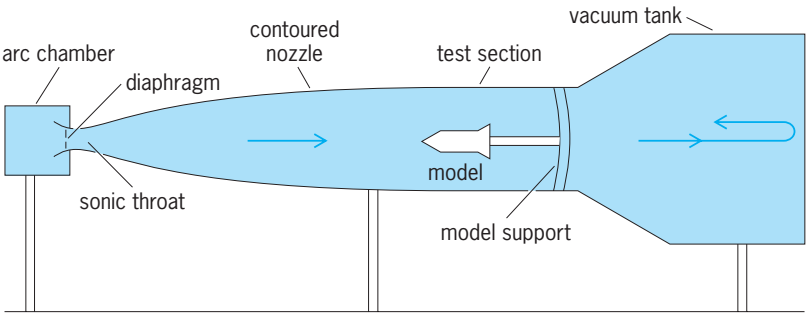


Fig. 9. Hypervelocity hotshot wind tunnel.

Complete duplication of all flight conditions in a hypervelocity tunnel is extremely difficult, if not impossible. For example, complete simulation of flight at 18,000 ft/s (5500 m/s) at 150,000-ft (45,000-m) altitude would require a tunnel supply temperature of 9000 K (16,000°F) and a supply pressure of 30,000 atm (450,000 lb/in.^{2} absolute or 3 gigapascals). Operation at such a high temperature is difficult, and operation at such a high pressure is not feasible at present. However, increasing the flight altitude to 300,000 ft (90,000 m) at the same velocity reduces the supply pressure required to a value that allows duplication to be achieved. Interpretation of test results is difficult, however, because the test gas is not in equilibrium owing to its rapid expansion from a highly ionized and dissociated state in the reservoir.

The major problem in designing a hypervelocity tunnel is the prevention of structural damage due to heating. Even though testing times may last only a fraction of a second, the high supply temperature (combined with the high pressure) results in a large heat flux, which reaches a maximum at the sonic throat. Damage at this point can result in large flow distortions, as may be inferred from the following example: A 0.025-in.-diameter (0.64-mm) sonic throat is required to establish a Mach 25 flow in a 115-cm-diameter (45-in.) test section. A change in throat diameter of only 0.002 in. (0.05 mm) would change the test-section Mach number by approximately 1.0.

The major types of currently operating hypervelocity tunnels are described below.

Hotshot tunnel. Mach numbers of 10–27 with testing times of 10–100 milliseconds can be obtained with this tunnel. The principle of operation (**Fig. 9**) is to prepressurize the arc chamber to 500–10,000 lb/in.^{2} abs (3.4–70 MPa) and evacuate the rest of the tunnel to a pressure of about 1 micrometer of mercury (approximately 10^{-6} atm or 0.1 pascal). The two chambers are separated by a plastic or a metal diaphragm. Electrical energy (capacitance or inductance storage of up to 10^8 joules) is then discharged into the arc chamber. This energy increases the air temperature and pressure, resulting in rupture of the diaphragm. The heated gas is then accelerated in the nozzle to provide hypervelocity flow. The test is usually terminated by a quick-acting dump valve which releases the stagnation chamber gas through a large port rather than requiring all the hot gases to pass through the sonic throat. This minimizes damage to both the throat and the interior of the arc chamber. The time at which this dump valve is actuated (that is, the effective run time) is influenced by a number of factors: (1) the run time necessary to obtain the desired data; (2) the time at which flow will "break down" at the model due to insufficient pressure ratio (resulting from either too large a decay in stagnation pressure or too large a buildup in vacuum tank backpressure); or (3) the maximum time the arc chamber or sonic throat can withstand the high heat-transfer rates.

Hotshot tunnels have been designed to operate at arc-chamber (supply) pressures and temperatures as high as 100,000 lb/in.^{2} abs (700 MPa) and 10,000 K (18,000°F), respectively. However, melting or oxidation of the tungsten throats (diameters of 0.020–0.40 in. or 0.5–10 mm) has, in general, limited useful arc-chamber pressures and temperatures to approximately 30,000 lb/in.^{2} abs (200 MPa) and 4000 K (7000°F). Nitrogen, rather than air, is used as a test medium in some tunnels to avoid oxidation of the tungsten throats. Other problems are the flow contamination from solid particles (throat or electrode melting) or foreign gases (pyrolysis of plastic diaphragms and insulators) and depletion of the test gas oxygen (when air is used).

Hotshot tunnels are used for heat transfer, force, dynamic stability, and pressure model tests; data are recorded on short-response-time oscillographs or on magnetic tape. Schlieren or shadowgraph photographs and high-speed motion pictures of the luminous airflow about the model provide data on flow formation, shock waves, and flow separation.

Shock tunnel. Test Mach numbers of 6 to 25 (velocities of 4000–15,000 ft/s or 1200–4500 m/s) with testing times from 0.5 to 4 ms can be obtained in this type of facility. The major sections of a shock tunnel (**Fig. 10**) consist of the driver, a constant-area tube or driven section, and a nozzle and dump tank. A high-velocity shock wave is produced by the rupture of the diaphragm separating the high-pressure driver section, containing helium, hydrogen, or a combustible mixture, from the low-pressure driven section, containing air or nitrogen. As the shock advances, it heats and accelerates the working gas behind it in the driven tube, and a region of uniform conditions is created between the shock wave and the gas interface (the contact surface separating the driver gas and the driven gas). The shock is reflected at the nozzle throat and passes back up the driven section, and the high-pressure high-temperature driven gas ruptures the throat diaphragm and accelerates to a high Mach number in the nozzle.

In a conventional shock tunnel, the reflected shock is again reflected at the gas interface and returns downstream; the run ends when the shock arrives at the nozzle throat for the second time. Run times of 4 ms are obtained by use of the "tailored interface" technique, which consists of fixing the properties of the gas on both sides of the interface so that the primary reflected shock wave produces equal pressures and velocities on each side of the interface. When this condition is met, the shock passes through the interface without reflection.

Working-gas (supply) pressures are normally in the range of 6000–20,000 lb/in.^{2} abs (40–140 MPa), while gas (supply) temperatures up to 10,000 K (17,5000°F) are achievable. Initial pressure in the vacuum dump tank is about 1–5 μm of mercury (0.1–0.7–0.7 Pa).

The main advantage of the shock tunnel over the hotshot tunnel is that the test medium is less contaminated with metallic particles, although some difficulty has been experienced with driver-gas contamination due to mixing at the interface. The main disadvantage of the shock tunnel is the shorter run time.

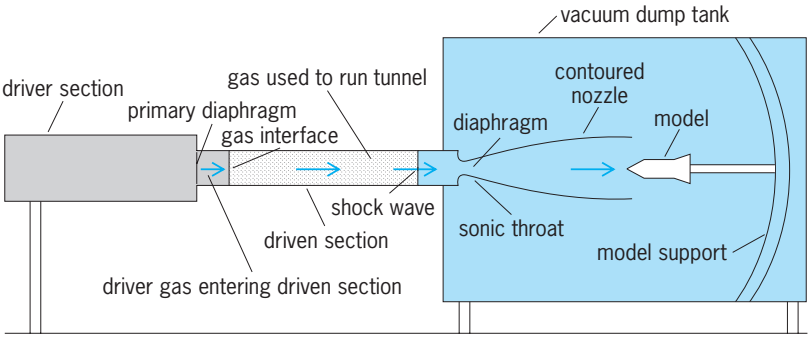


Fig. 10. Hypervelocity shock wind tunnel.

Instrumentation with response times of less than a millisecond has been developed for measurement of heat transfer, pressure distribution, and aerodynamic forces on models. A particular problem in measuring aerodynamic forces on models arises from the necessity of making the measurements while the model is still vibrating from the impact absorbed when the starting shock passed over it. This vibration produces inertial forces which must be measured with accelerometers and accounted for in the force balance data reduction. *See* SHOCK TUBE.

Graphite heater blowdown tunnel. This type of tunnel has not yet earned a nickname, so it is labeled in a very generic way. Test Mach numbers up to 14 with testing times from 0.25 to 2.5 s can be obtained with this relatively new design of hypervelocity tunnel. A Mach 18 facility has been designed but not yet built. In operation, nitrogen at room temperature is used to fill the heater (**Fig. 11**) to a pressure on the order of one-fourth of the desired test pressure. Electrical power (up to a megawatt) is then applied to a graphite heater element which heats the gas at constant volume, increasing its pressure to the desired stagnation pressure. The nitrogen is confined in the heater by a double diaphragm which is ruptured by overpressuring the interdiaphragm volume to begin the run. The decay in stagnation pressure associated with the hotshot tunnel is eliminated by introduction of controlled pressure cold gas into the bottom of the heater to push the hot test gas up, in a pistonlike fashion, out of the heater. The time over

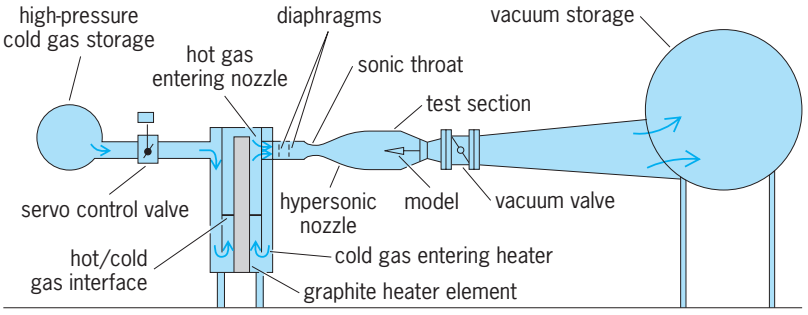


Fig. 11. Hypervelocity graphite heater blowdown tunnel.

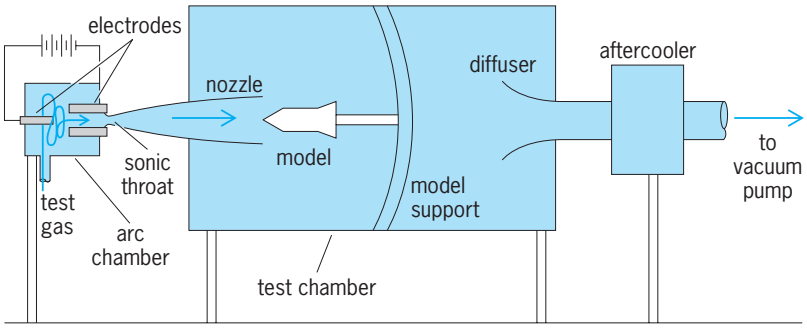


Fig. 12. Hypervelocity plasma-jet wind tunnel using dc power.

which valid data may be taken is terminated by arrival of the hot/cold gas interface at the nozzle. Control of the cold nitrogen pressure is maintained by very fast-acting electrohydraulic servo-driven control valves. Maximum supply pressures of 20,000 lb/in.^{2} abs (140 MPa) and supply temperatures of 3200°R (1800 K or 2750°F) have been obtained in this type of facility.

The main advantage of this type of tunnel over the shock tunnel is the relatively long run times. During the test time available in this facility, it is possible to make a complete high speed angle-of-attack sweep from 0 to 20 or 25°, obtaining heat transfer, pressure, or aerodynamic force data over that entire angle-of-attack range. In either a shock tunnel or a hotshot, the model is preset to given angle of attack, and data are obtained only at that condition during the run. Its main disadvantage lies in its complexity and resultant high cost.

Plasma-jet tunnel. This tunnel has the capability of developing the highest temperature (approximately 20,000 K or 35,000°F) and the longest run time (several minutes) of any hypervelocity tunnel (**Fig. 12**). It is arc-heated and utilizes either dc or ac power. The arc heater consists of a water-cooled cathode, an injection (or vortex) chamber, and a water-cooled hollow anode.

The function of the injection chamber is the generation of a strong vortex to stabilize the arc in the center of the heater. The vortex is created by the injection of high-pressure air (or other test gases) in a tangential direction at the outer edge of the vortex chamber. After injection, the test gas flows to the center of the injection chamber, through the hollow anode, and out the nozzle throat. Since the angular momentum of the test gas will be conserved as it flows to the center of the chamber, its tangential velocity will increase and its density will decrease. This effect creates a pronounced radial density gradient in the vortex chamber and hollow anode, causing the arc to remain along the center line, where the density is lowest.

A magnetic field is used to rotate the arc so that the arc-attachment points at the cathode and anode do not become overheated. Further cooling of the anode is provided by the flow of test gas, since the swirling motion of the flow tends to keep cooler gas along the anode surface. A secondary benefit of this phenomenon is an increased heater efficiency because of reduced heat transfer to the anode and nozzle throat. Overall efficiencies as high as 60–70% can be obtained with a gas-stabilized arc heater.

The hypervelocity facility is used primarily for heat-transfer tests, reentry-shape ablation studies, and magnetoaerodynamic studies. Relatively high velocities (up to 20,000 ft/s or 6000 m/s) are achievable, but few plasma tunnels have been proposed or built for obtaining Mach numbers greater than 10. Maximum operating supply pressure is about 3500 lb/in.^{2} abs (24 MPa). At high pressures the power requirements are greatly increased and arc stabilization is difficult. Plasma-jet units with power requirements of 1–60 MW are in operation or under construction; units with power requirements of several hundred megawatts have been proposed.

Disadvantages are electrode contamination and large nonuniformities in the test-gas stream, high power requirements, and operating pressure limitations. Instrumentation consists of high-speed photographic equipment, calorimeter, pitot and mass flow probes, microwave and electron beam apparatus, and a spectrograph.

Other hypervelocity wind tunnels are the gun tunnel, shock tube, expansion tube, and wave superheater tunnel. Randall C. Maydew; Donald D. McBride

Instrumentation

Specifically designed mechanical, electrical, and optical devices are used to measure effects of air flow across a model. Wind tunnel testing has always required instrumentation that is rugged, reliable, and accurate. Increased costs for energy and worker-power has brought demands for shorter run time and higher productivity. To fulfill these needs, modern instrumentation has been developed which can acquire accurate data quickly and provide information directly to the research engineer during the test. Measurements common to most tunnels are pressure, temperature, turbulence, and flow direction.

Pressure measurement. The earliest method used to measure a large number of pressures in tunnel tests was with multitube liquid manometer boards. A photograph of the board was taken at each special test configuration. Later, the height of each column was measured on the photo to provide the data.

Many techniques have since been developed to provide direct conversion of pressure to electrical signals. Individual strain gage, capacitance, or force balance transducers may be used, but when several hundred pressures are read this method is expensive and requires much set-up time. The mechanically scanned multiport sampling valve which switches, in sequence, 48 pressure tubes to a single transducer is often used. Since, for an accurate reading, time must be allowed for the pressure to settle after it is switched, the maximum reading rate is about 20 pressures per second.

An electronically scanned pressure (ESP) measuring system allows pressure sensors to be sampled at 10 kHz. The heart of the system is the electronically scanned pressure module (**Fig. 13**) which contains 32 silicon pressure transducers, an electronic

multiplexer, and a calibration valve in a volume less than 4.2 in.^{3} (69 cm^3). The modules may be mounted within a tunnel model to avoid the slow repsonse caused by long pneumatic tube lengths. Because all transducers are periodically calibrated in place against an accurate pressure standard, the system accuracy is within ± 0.15% of full scale.

Some types of testing, such as engine inlet dynamics, require measurement of pressure fluctuation up to 20 kHz. Rakes containing as many as forty 0.08-in.-diameter (2-mm) semiconductor strain gage transducers are used for this purpose. *See* PRESSURE MEASUREMENT.

Temperature measurement. Most temperature measurements in tunnel tests are made with thermocouples, thermistors, and platinum resistance temperature sensors. The airstream temperature measured is the stagnation or total temperature. The static temperature is computed from this measurement.

Heat transfer measurements may be made by constructing a thin-wall tunnel model with thermocouples closely spaced on the inside surface of the wall. Individual heat transfer gages can be built by using a thin copper disk with a fine thermocouple attached, mounted in an insulated holder. The gage is installed with the sensing disk flush with the outside model surface. A heat transfer gage used in shock tunnels is constructed as a thin-film resistance thermometer. Using a 0.1-μm-thick platinum film deposited on a Pyrex glass substrate, a response time of 10^{-7} s is obtained. *See* TEMPERATURE MEASUREMENT.

Turbulence measurement. The level of velocity fluctuations is important in tunnel testing because it influences the point on a model at which the boundary layer changes from laminar to turbulent. This point of transition affects the aerodynamic drag forces on the model. Two instruments used to measure turbulence are the thermal anemometer and the laser velocimeter. *See* BOUNDARY-LAYER FLOW; TURBULENT FLOW.

The thermal anemometer is used to obtain instantaneous velocity and flow angle measurements. The

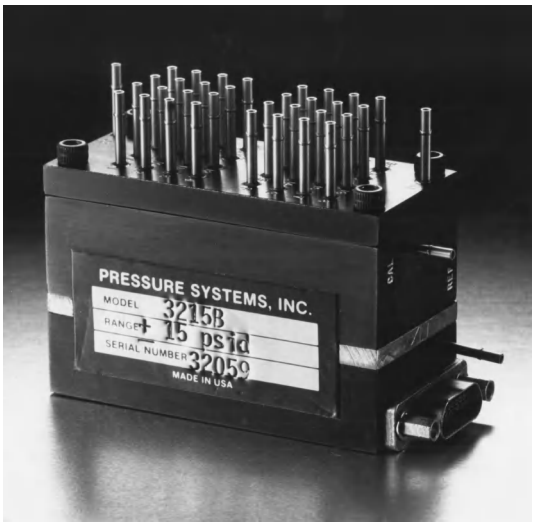


Fig. 13. A 32-transducer electronically scanned pressure module. (*Pressure Systems, Inc.*)

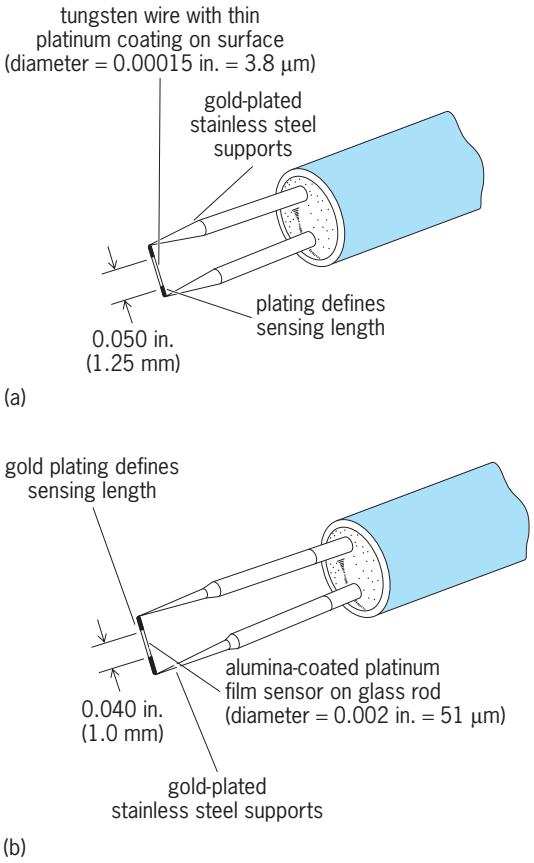


Fig. 14. Thermal anemometers with (*a*) hot-wire sensor and (*b*) cyclindrical hot-film sensor. (*TSI Inc.*)

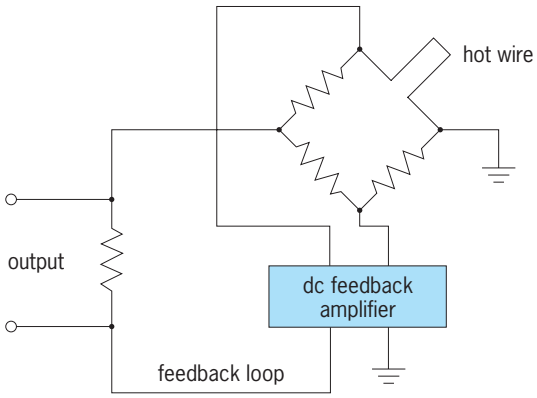


Fig. 15. Constant-temperature thermal anemometer circuit.

sensor contains a fine wire (**Fig. 14*a***) or a conducting film on a ceramic substrate (Fig. 14*b*) which is electrically heated to a temperature greater than the surrounding air. Since the amount of heat conducted away from the sensor depends on the local air velocity, its temperature will change accordingly. The resistance of the sensor varies with its temperature, which allows its use in a Wheatstone bridge circuit to provide a signal related to flow. *See* ANEMOMETER.

The hot wire or film by itself cannot respond to changes in fluid velocity at frequencies above 500 Hz. With electronic compensation, this response can be increased to over 1 MHz. With the constant temperature circuit, shown in **Fig. 15**, a servoamplifier is

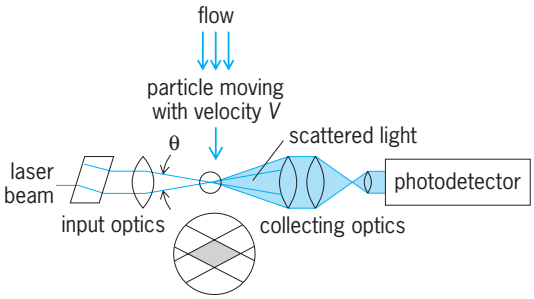


Fig. 16. Configuration of laser velocimeter.

used which senses an unbalance in the bridge due to a change in sensor resistance and feeds back a current to the bridge to restore balance. In this way the wire is held at a constant instantaneous resistance, and thus a constant temperature. Therefore, no thermal lag occurs, and the feedback current is a direct measure of the turbulence.

The laser velocimeter (LV) is an optical technique that provides remote measurement of mean velocity and turbulence in the flow field of a tunnel model. Since no mechanical probe is used, there is no disturbance to the flow. As shown in **Fig. 16**, a laser beam is optically split into two equal-intensity beams and focused to a point in the flow field. At this point, the light waves from each beam interfere constructively and destructively, forming a set of interference fringes. Extremely small aerosol particles (about 1-μm diameter) are injected upstream of the focal point. As each of these particles follows the airstream through the bright and dark fringes, the light that it scatters will vary in intensity. The scattered light is imaged onto a photodetector which produces an electrical signal proportional to the light intensity. The component of velocity normal to the optical axis in the plane of the two input laser beams can be determined from the signal frequency with the equation below, where V is the component velocity

$$V = \frac{\lambda f}{2 \sin\left(\frac{\theta}{2}\right)}$$

(m/s), f is the signal frequency (Hz), λ is the wavelength of the laser beams (m), and θ is the angle between the two laser beams. *See* LASER.

By using a two-color laser beam and additional optics, velocities in two directions can be measured simultaneously. A two-color laser velocimeter with blue and green beams is shown in **Fig. 17**.

A wide variety of flow visualization techniques are available for observing turbulent flow. Tufts attached to the model show flow patterns near the surface. China clay, lampblack, or fluorescent dyes supsended in viscous oil are used for studying flow in the boundary layer. Sublimation of fluorene or azobenzene is used to study extremely thin boundary layers. Schlieren photos allow the flow to be visualized in supersonic airstreams.

A method developed to visualize vortex patterns is called the laser sheet. A laser beam is fanned out into a thin vertical sheet and projected normal to the tunnel axis. A fine mist injected into the tunnel scatters the projected light and forms a visible screen. Any disturbance of the uniform flow field, such as that produced by vortices, appears as dark holes within the vapor screen.

Direction probes. The flow direction and the degree to which the flow is parallel to the tunnel walls is vital in the analysis of pressure and forces on a test model in a wind tunnel. Several devices are used, depending on the flow velocity (**Fig. 18**).

Yaw sphere. In subsonic flow the pressure distribution on the surface of a sphere can be used to

Fig. 17. Laser velocimeter used for turboprop flowfield analysis in wind tunnel at NASA Lewis Research Center.

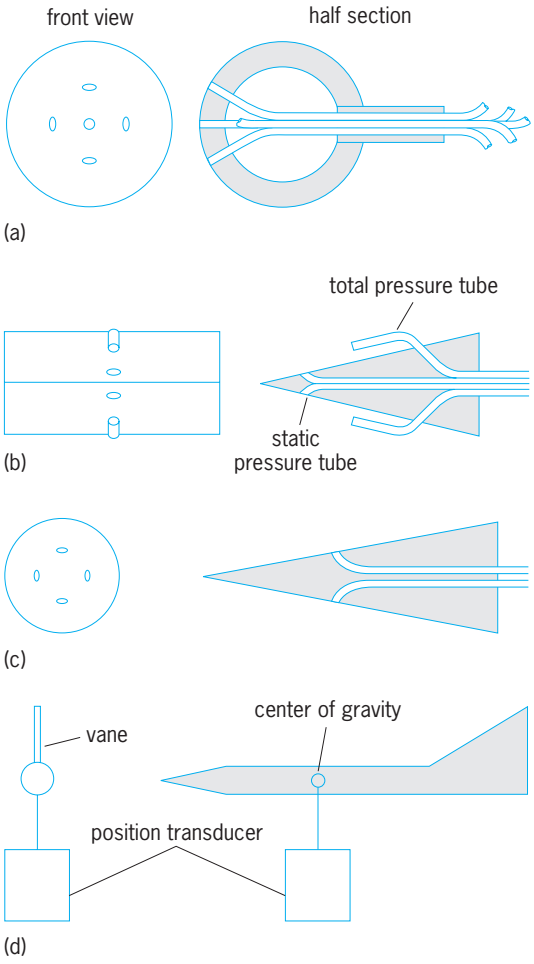


Fig. 18. Direction-measuring instruments. (*a*) Yaw sphere. (*b*) Wedge. (*c*) Cone. (*d*) Vane.

determine the flow direction of the air passing over the sphere. In a typical probe, yaw angle and pitch angle can be determined simultaneously. The ratio of the difference between opposite static pressures to the average of these pressures should be calibrated against the flow angles before the probe is used in a tunnel.

Wedge. In supersonic flow the sphere is not usable; instead, the wedge serves in its classical use as a means to determine the velocity of a supersonic stream and as an angle-of-flow meter. The wedge is a two-dimensional device, so that pitch and yaw direction measurements must be made separately. When the wedge is used only for directional measurement, the total-pressure tubes are not needed if the Mach number is known. The angle of flow can be obtained from tabulated data for wedges in supersonic flow. The wedge with static-pressure tubes can be used only to indicate subsonic flow angles, but first must be calibrated.

Cone. Similar to the wedge, the cone is usable in both supersonic and high subsonic flow fields to determine flow angles. The cone can be used to measure yaw and pitch angles simultaneously. However, like the sphere, it must be calibrated.

Vane. When space is not restricting, the mechanical vane can give direct-reading angles of yaw or pitch without detailed computations. The vane is mounted at its center of gravity on a calibrated position transducer and is read remotely. Daniel J. Shramo

Bibliography. J. D. Anderson, Jr., *Hypersonic and High Temperature Gas Dynamics*, rev. ed., 2000; J. B. Barlow, W. H. Rae, Jr., and A. Pope, *Low-Speed Wind Tunnel Testing*, 3d ed., 1999; F. Durst, A. Melling, and J. H. Whitelaw, *Principles and Practices of Laser-Doppler Anemometry*, 2d ed., 1981; N. Isyumov (ed.), *Wind Tunnel Studies of Buildings and Structures*, rev. ed., 1998; A. Pope and K. L. Goin, *High-Speed Wind Tunnel Testing*, 1965, reprint 1978.

Windings in electric machinery

Windings can be classified in two groups: armature windings and field windings. The armature winding is the main current-carrying winding in which the electromotive force (emf) or counter-emf of rotation is induced. The current in the armature winding is known as the armature current. The field winding produces the magnetic field in the machine. The current in the field winding is known as the field or exciting current. *See* ELECTRIC ROTATING MACHINERY; GENERATOR; MOTOR.

The location of the winding depends upon the type of machine. The armature windings of dc motors and generators are located on the rotor, since they must operate in conjunction with the commutator, and the field windings are mounted on stator field poles. *See* DIRECT-CURRENT GENERATOR; DIRECT-CURRENT MOTOR.

Alternating-current synchronous motors and generators are normally constructed with the armature winding on the stator and the field winding on the rotor. There is no clear distinction between the armature and field windings of ac induction motors or generators. One winding may carry the main current of the machine and also establish the magnetic field. It is customary to use the terms stator

Fig. 1. Salient-pole field winding on rotor of synchronous motor. (*Allis-Chalmers*)

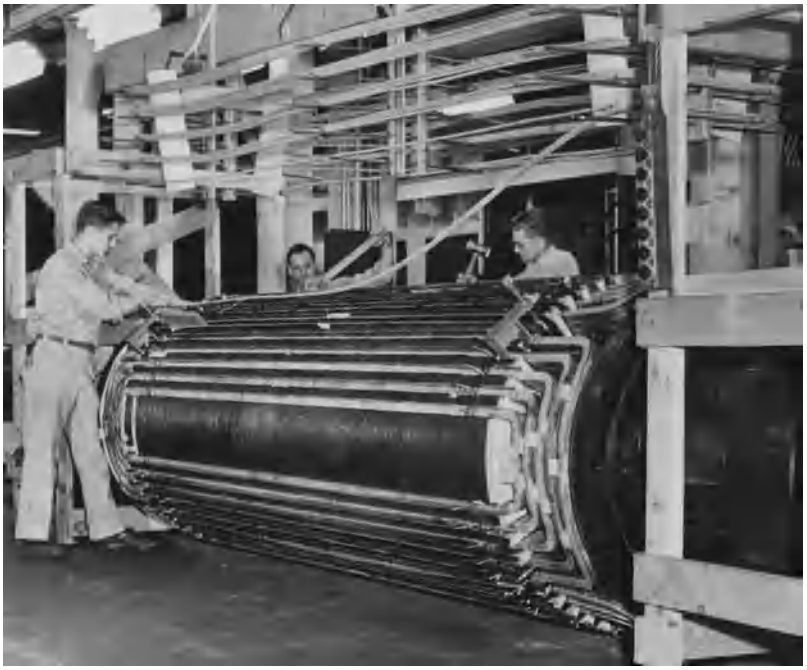


Fig. 2. Winding field on cylindrical rotor. (*National Electric Coil Co.*)

winding and rotor winding to identify induction motor windings. The word armature, when used with induction motors, applies to the winding connected to the power source (usually the stator). *See* ALTERNATING-CURRENT GENERATOR; ALTERNATING-CURRENT MOTOR; SYNCHRONOUS MOTOR.

Field windings. Field windings produce a magnetic pole fixed in space with respect to the magnetic structure on which they are mounted. If salient-pole construction is employed, the winding turns are concentrated around the pole core (**Fig. 1**). If cylindrical construction is employed, the field winding is constructed of elongated concentric loops embedded in slots cut in the surface of the field structure. In the case of dc motors and generators with compensating windings, both forms of construction may be found in one machine.

Concentrated field windings for salient-pole ac synchronous machines and for shunt fields of dc and series ac machines are form-wound of many turns of insulated copper wire. The completed coil is taped and impregnated to hold its shape and to fit around the pole core. Series field coils are normally concentrated. They are wound of larger wire, or of insulated copper strip, in order to carry the armature current without overheating.

Field coils for cylindrical construction are made of insulated rectangular copper or aluminum bars (**Fig. 2**). Individual coils are form-wound with several turns per coil. The dimensions of the coils are selected to provide a suitable space distribution of field flux. The coil is held in the slots by nonmagnetic wedges. The coil ends are held in place against mechanical forces by nonmagnetic bands.

Rotor field coils are connected to the external circuit through slip rings, on which carbon brushes rest to conduct the exciting current. *See* SLIP RINGS.

Armature windings. These windings carry ac current. They take many forms, depending upon the type and capacity of the machine and specific design requirements. Armature windings have an active portion, which lies in slots in the magnetic circuit, and end turns and end connections, which are external to the airgap. The treatment of the end connection determines the appearance and operating characteristics of the winding. Usually the end turns are formed to make diamond-shaped coils for ease in inserting and bracing the windings. A coil may have a single turn, or it may have many turns. Each turn may have a single strand, or it may have several strands of copper electrically in parallel. A coil has two coil sides, which lie in slots approximately a pole pitch apart for dc machines, whereas most ac machines have short-pitch windings in which the coil pitch is less than 180 electrical degrees. The pitch of a coil is found by taking the quotient of the slots separating the coil sides and the slots in 180 electrical degrees. Windings may be single layer, one coil side per slot, or double layer, two coil sides per slot, in which case one side of each coil will rest in the bottom half of its slot and the other coil side at the top of its slot.

Direct-current armature windings. These may be lap, or parallel, windings, or they may be wave, or series, windings. **Figure 3** shows a four-pole lap dc armature winding and a four-pole wave dc aramature winding. The two positive brushes of the lap winding would be connected together and the two negative brushes would be connected together. The lap and wave windings have the following general properties:

1. Coil ends of lap windings are fastened to adjacent commutator bars. Coil ends of wave windings are fastened to commutator bars approximately two pole pitches apart.
2. The lap winding has as many parallel paths between the line terminals as there are poles, and the conductors of a path lie under adjacent poles. The wave winding has two parallel paths between the line terminals regardless of the number of poles, and conductors of either path lie under all poles.
3. The lap winding has as many sets of brushes as there are numbers of poles, whereas wave windings need only two brush sets. Usually, however, the

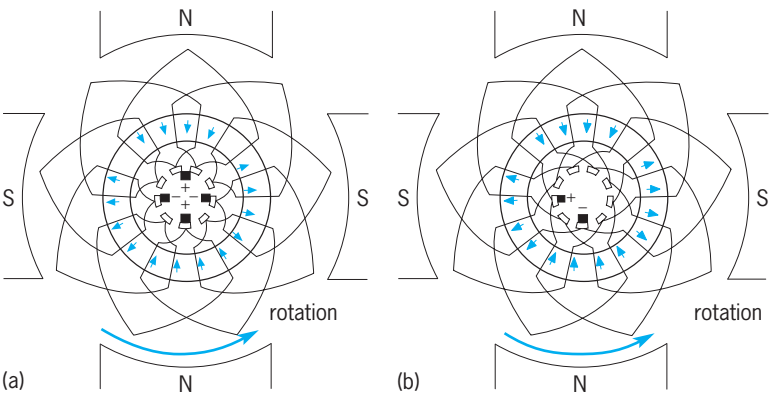


Fig. 3. Armature windings. (*a*) Simple four-pole lap winding. (*b*) Simple four-pole wave winding.

machine carries as many brush sets as there are poles to provide greater current-carrying capacity for the commutator. Brush positions are such that the coils are short-circuited only when they are in the position of minimum induced voltage.

Lap windings are adapted to high-current machines because they may have more than two parallel paths, whereas the wave windings are adapted to small-capacity machines and high-voltage machines because of the series connection of the coils.

Alternating-current armature windings. These may be single or polyphase, full or fractional pitch, Y-connected or delta-connected. Practically all except small-capacity ac machines have three-phase windings, with coils distributed around the entire armature periphery for better utilization of space and material in the machine. Winding factors are used to evaluate the performance of a winding. The generated rms voltage per phase is $E = 4.44\, k_d k_p f N_{pb} \Phi$, where k_d is the distribution factor, k_p the pitch factor, f the frequency in hertz, N_{pb} the total turns in series per phase, and Φ the flux per pole in webers per square meter.

The distribution factor k_d of a winding gives the ratio of the voltage output of a distributed winding to the voltage output of a concentrated winding. This is shown by Eq. (1), where n is number of slots over

$$k_d = \frac{\sin(n\gamma/2)}{n \sin(\gamma/2)} \tag{1}$$

which one phase is distributed and γ is electrical angle between slots. Fractional-pitch windings are employed to reduce the magnitude of harmonics, to improve wave shape, to save copper, and to permit end connections between coils to be made more easily.

The pitch factor k_p of a winding gives the ratio of the voltage output of a fractional-pitch coil to the voltage output of a full-pitch coil. This is shown by Eq. (2), where ρ is electrical angle between coil sides.

$$k_p = \cos\frac{\pi - \rho}{2} \tag{2}$$

Since the angle γ and the angle ρ in the equations for k_d and k_p change with frequency, they may be selected to eliminate or minimize certain harmonic frequencies.

Figure 4 shows the active conductors of a distributed, three-phase, two-pole, fractional-pitch, armature winding wherein a, b, and c are conductors of three phases displaced 120 electrical degrees apart. Coil a_1, $-a_1$ spans $\frac{5}{6}$ of a pole pitch, or 150 electrical degrees.

Stator and wound-rotor windings of polyphase induction motors conform to the above descriptions. Single-phase induction motor stators may be wound in the same manner, except for the number of phases; or they may be formed of several concentric loops of wire. The windings are usually form-wound from one continuous length of wire, but are not taped. This facilitates the installation of the winding in small, partly closed slots. For additional information on these and squirrel-cage windings *see* INDUCTION MOTOR

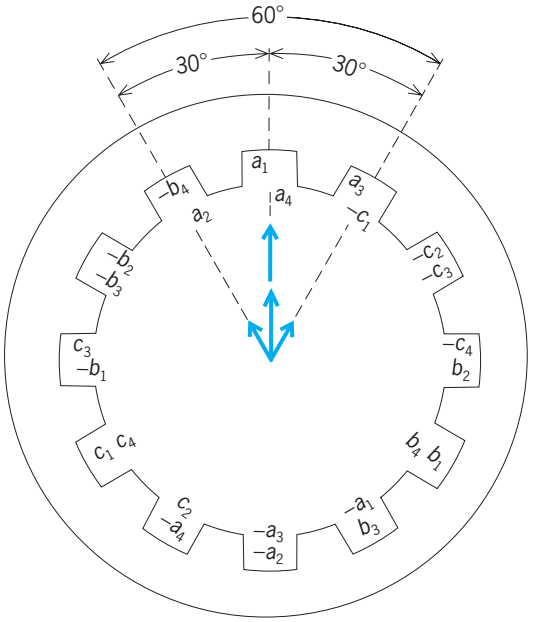


Fig. 4. Distributed, three-phase, two-pole, fractional-pitch armature winding. The voltage vector diagram is given. (*After A. E. Fitzgerald and C. Kingsley, Electric Machinery, McGraw-Hill, 1952*)

Arthur R. Eckels

Bibliography. R. T. Anderson et al., *Large Rotating Machine Windings*, 1969, reprint 1981; A. E. Fitzgerald, C. Kingsley, and S. D. Umans, *Electric Machinery*, 5th ed., 1990; M. Liwschitz-Garik and C. Gentilini, *Winding Alternating Current Machines*, 1950, reprint 1975; S. A. Nasar, *Electrical Machines and Power Systems*, 1995.

Wine

Technically, the juice of any fruit that has been subjected to an alcoholic fermentation. However, both in general usage and by legal definition the term wine is defined as a beverage produced by the alcoholic fermentation of the juice of grapes.

Grapes possess two attributes that contribute to their preeminence in wine production. The juice of fresh ripe grapes is mildly acidic (pH 3.0–3.5) and contains about 21–25% fermentable sugar (mostly glucose and fructose). These properties favor the multiplication of naturally occurring yeasts which, under anaerobic conditions, carry out an alcoholic fermentation that yields ethanol and carbon dioxide as the major end products. The final result is a beverage with sufficient ethanol concentration (usually 12.5% by volume) to render it relatively stable if it is kept in sealed containers, and sufficient in tartness and flavor to be interesting to the palate. Other fruits yield stable and flavorful fermented beverages only by considerable artifice, usually involving supplementation of their juice with fermentable sugar

and with acid. *See* ETHYL ALCOHOL; FRUCTOSE; GLUCOSE.

Grape species. Most wines are made from various cultivars of the classic European wine grape species, *Vitis vinifera*. However, in the United States east of the Mississippi, wine is also made from varieties of two native American grape species, *V. labrusca* (varieties descended from the wild "Fox Grape" still found in New England forests) and, to a lesser extent, *V. rotundifolia*. The juice of *V. labrusca* grape varieties is often too acid for production of salable wine. In order to derive a palatable wine from these grapes, their musts (that is, the juice) may be legally diluted with up to 35% by volume of water or sugar solution, prior to fermentation (without disclosure of the fact to the consumer). This process is euphemistically termed amelioration. In other countries the admixture of water or a sugar solution to musts in the production of wine is forbidden; however, musts may be legally sugared in Germany, provided the products are labeled *Verbessert*. Increasingly, *V. vinifera–V. labrusca* hybrids have been planted in many wine-growing areas of the eastern United States. Wines from the best of these crosses have little of the methyl anthranilate odor associated with wines from *V. labrusca* varieties. *See* GRAPE.

Wine types. Wines are classified on the basis of the presence or absence of carbon dioxide, on alcohol content, color, and relative sweetness.

Alcohol content. Wines whose dissolved carbon dioxide content is insufficient to produce long-lasting effervescence when served are called still wines, and are divided on the basis of their alcohol content into three general categories: table wines, fortified wines, and so-called soft or light wines. Table wines, containing 9–14% alcohol by volume, are usually used to accompany the main courses of a meal; their alcohol content derives directly from the sugar present in the original grape juice. Fortified wines have alcohol concentrations up to 20% by volume, achieved by adding brandy (spirits distilled from wine) at some stage of their production. They are sometimes known as dessert wines because such wines are often served at the end of a meal. Established as a class in the early 1980s, light wines, which contain only about 6–7% alcohol, are often prepared by removing some of the alcohol from table wines by vacuum distillation. However, some of the flavor constituents are also removed by this process, and the final product is essentially similar to a table wine which has been diluted with water. Reverse osmosis technology is used to produce other liquids legally classified as wine. In this procedure, standard wine is routed, under pressure, through a filtering chamber housing membranes permeable to substances having a nominal molecular weight of less than 500, thereby separating the standard wine into a permeate and a retentate, the latter containing reduced levels of alcohol and water and legally sold as wine.

Color. White wines (usually pale straw in color) are most often made from light-skinned grape varieties, but can also be made from dark-skinned varieties inasmuch as the juice of most wine grapes, light-skinned or dark-skinned, is unpigmented. When white wines are made from dark-skinned grapes, the juice is separated from the skins immediately after the grapes are crushed, before the anthocyanin pigments of the skins begin to dissolve. Relatively short contact (2–6 h) of colorless juice with the skins of crushed dark-skinned grapes allows extraction of sufficient color to yield final products that are various shades of pink. Technically, such products are rosé wines, but they may be labeled as white wines (for example, White Zinfandel) or even as gray (as in *vin gris*).

In red wine production the skins of crushed, dark-skinned grapes are allowed to remain in contact with the grape juice for longer periods of time during the course of the alcoholic fermentation, and the pigments of the skins become increasingly soluble as the fermentation proceeds; the maximum solubility of the pigments occurs when the alcohol concentration of the fermenting juice reaches about 6% by volume. However, long contact of skins and fermenting juice does not necessarily deepen the color of the final wine: when the cells of the grape skins die as the fermentation proceeds, they themselves become adsorbers of the very pigments that were extracted from them earlier. Thus, for maximum color to remain in the wine it may be necessary to remove the fermenting juice from the skins before the fermentation has gone to completion. Removal of the skins from a red wine fermentation before the fermentation is complete is also sometimes done for other reasons. If only a light red color is desired in the final product, the juice is separated from the skins early. This happens also when appreciable tannin extraction (which roughly parallels color extraction) must be avoided, as when wine is specifically being made for early consumption.

Relative sweetness. Purely on a compositional basis, wines are classified as dry or sweet, terms referring to the relative sugar content of the wine. Products whose initial sugar content has been completely, or almost completely, fermented away are dry wines. Sweet wines contain residual sugar and have a recognizably sweet taste.

The occurrence of residual sugar in wines may be achieved naturally or by technical means. For example, if a wine is made from partly desiccated grapes (that is, either raisined, or affected by the mold *Botrytis cinerea* and partly dehydrated), the must may contain more sugar than can be fermented to ethanol by wine yeasts before the ethanol concentration becomes inhibitory (most wine yeasts are intolerant of alcohol concentrations above 14%, but some can tolerate up to 16%). Thus wines made from such musts will, naturally, contain residual sugar. Examples are the Sauternes of France, the *Beerenauslesen* and *Trockenbeerenauslesen* of Germany, the *Vendange tardive* wines of Alsace, the *vini passiti* of Italy, and the Late Harvest wines of California.

Other wines are made to contain residual sugar by the technique of stopping the fermentation of the sugar in the musts of grapes of normal ripeness (that is, grapes neither raisined nor botrytised) by

cooling the fermenting juice and continuing the fermentation under pressure once the desired residual sugar concentration has been reached (both environmental factors acting to halt yeast activity), then filtering the partially fermented product free of yeast cells. This technique is widely used in the production of slightly sweet German white wines. Halting fermentation can also be achieved by adding brandy to a fermenting must before all the sugar has been fermented away. This is the technique used in the production of Porto, in Portugal, and of Porto-type dessert wines, red or white, elsewhere.

Finally, to impart requisite sweetness, the addition of unfermented grape juice (*muté* or *Restsüsse*) preserved with sulfur dioxide to wines that have been fermented to dryness is also sometimes employed. Unfortunately, this technique yields products tasting of grape juice and containing relatively high concentrations of sulfur dioxide, a substance that, in its free state, is not only organoleptically unpleasant but also harmful to asthmatics and other sensitive individuals.

Wine production. Winemaking operations vary widely with the type of wine being produced. In the production of table wines the grapes, harvested when they have reached the desired sugar and acid levels, are normally introduced into a crusher-stemmer, a mechanical device that crushes the berries but not the seeds, and removes the stems. Removing the stems is a standard but not universal practice. Some fine red wines of Burgundy are fermented with the stems in order that stem tannins may be extracted into the wine. The skins of the grape variety Pinot noir from which these wines are made are themselves low in tannins which, as reducing agents, help protect wines from oxidation and give red wines longevity. The stems are also not removed prior to fermentation in the production of the red wines of the Beaujolais. In this region of France, large fermenters are filled with whole grape bunches and sealed almost completely, with only a small opening being provided at the top of the vats for gas escape. The weight of the grapes at the top crushes some of the berries at the bottom, and a fermentation of the crushed berries begins, releasing carbon dioxide which, being more dense than air, forms a layer above the fermenting juice. As the carbon dioxide layer rises and respiration of the cells of the grape skins above the crushed berries is prevented, the cells die and the grape skins burst, releasing more berry juice to be fermented. Eventually all the berries are crushed in this fashion, without the intervention of mechanical crushing machines. This process, known as carbonic maceration, results in red wines particularly fruity-vinous in fragrance.

When white table wines are being fermented, the skins and seeds (a mixture known as pomace) are separated from the juice, and the juice is drawn off either to stainless steel or glass-lined fermenters or, more traditionally, to wooden barrels. The barrels are positioned on their sides and filled to only 65% of their capacity to allow space for the frothing which inevitably occurs during the fermentation. Red wines are fermented in open vats, usually made of wood, but sometimes of stainless steel. As the fermentation of a red wine proceeds, the skins of the crushed grape berries fill with carbon dioxide and rise to the surface, forming a cap layer. This cap must be "punched down" or otherwise remixed with the fermenting juice periodically, so that proper extraction of skin constituents may continue and so that the heat released during the fermentation does not raise the temperature in the cap to the point at which the yeasts are killed, causing the fermentation to "stick" or halt. As the temperature of the fermenting liquid must be controlled, large fermentation tanks are often equipped with heat transfer coils; the temperature of the fermenters is usually maintained in the range 75–86°F (24–30°C).

After the initial fermentation the wine is cellared; red wines are separated from the pomace with a press. The now-alcoholic pomace can be used for distilling purposes (in Burgundy the distilled product is called *marc* and in Italy *grappa*), or it can be used as fertilizer. Red wines are almost always cellared in wooden barrels, in which the alcoholic fermentation is completed and the wine clears itself.

Because extractives leach into the wine from the wood, the wood type and barrel size—the latter determining the wood-to-wine ratio—are important factors in the taste of the final product. American white oak (particularly *Quercus alba,* although other American *Quercus* species are also utilized) is still widely used in the United States for aging table wines and American oak has been found to be the wood of choice for aging Jerez (that is, sherry) wines in Spain. However, European oak (*Q. robur* and *Q. sessilis*) from Limousin or Nevers, France, or Slovenia, Yugoslavia, is used in Europe and increasingly in the United States for aging fine wines from the best *V. vinifera* varieties. In California, large storage casks are often of redwood (*Sequoia sempervirens*), which is essentially taste-neutral in character.

As either red or white wines clear, the yeast cells, and the precipitates that form during fermentation (largely potassium bitartrate, the commercial cream of tartar), settle as a sediment (lees) at the bottom of the storage container. Periodically, and depending on the type of wine being produced, the new wine is drawn off (racked) from this sediment, and is transferred to clean containers. Aging is then continued, and many red wines are kept in wood for 2 years before being bottled. Some white wines are not aged in wood at all, but are merely racked and stored in glass-lined tanks until finally stabilized and bottled. Before bottling, wines may be fined by the addition of substances such as egg white, sturgeon-derived isin-glass, or bentonite clay to help settle the finely dispersed particles that have been suspended in the wine. Some wines are filtered, but tight filtering of red wines is controversial. The most famous (and most expensive) red table wines of the world are not clarified in this way, but are purposely allowed to go to market containing microscopic particles that act as nuclei for slow precipitate formation. Upon long aging, such wines throw sediments in the bottle,

consisting largely of oxidized tannins and anthocyanin coprecipitates. Some wines are cold-stabilized, that is, refrigerated for a short period of time, to bring down excess potassium bitartrate, which might precipitate in the bottle when the product is chilled. In the production of ordinary wine the use of ion-exchange resins for the prevention of potassium bitartrate precipitation is now common throughout the world, although it is not permitted in France. Some inexpensive and poorly made wines are pasteurized before bottling by a brief exposure to a temperature of 140°F (60°C) in the absence of air. This process, while preventing wine spoilage by microorganisms, causes hydroxymethyl furfural formation in the wine, and deleteriously alters the wine's flavor; hence it is avoided whenever possible.

At the time of bottling, white wines, in particular, are usually dosed with a small amount of sulfur dioxide to take advantage of its antiseptic and antioxidative effects. Total (that is, free and chemically bound) sulfur dioxide of 75–150 mg/liter, and free sulfur dioxide of 10–30 mg/liter are usual in well-made white wines. Well-made red table wines usually contain total sulfur dioxide of less than 0.0033 oz/quart (100 mg/liter), most of which is present in various chemically bound forms (sulfonates, for example) by the time the wine is sold.

Fermentation. Almost all naturally fermented table wines are produced by an alcoholic fermentation carried out principally by strains of the sporogenous yeasts *Saccharomyces cerevisiae* var. *ellipsoideus* or var. *oviformis*. The actual fermentation of the grape must, however, if it were allowed to occur completely in the absence of sulfur dioxide and as the result of the activities of the naturally present microscopic flora, would start with the multiplication of asporogenous yeasts such as *Kloeckera apiculata*. These so-called wild yeasts are usually not tolerant of alcohol concentrations in excess of 4% and so would be killed as the fermentation proceeded, to be succeeded then by various *Saccharomyces*. However, since the wild yeasts produce end products other than ethanol and carbon dioxide during their metabolism of glucose, and some of these products, such as acetic acid, are not desirable in wine, their multiplication in fermenting musts is usually discouraged. This is accomplished by the addition of a small amount of sulfur dioxide (in aqueous solution; or in the form of aqueous solutions of potassium bisulfite or metabisulfite) to the must as soon as the grapes are crushed, because most wild yeasts are inhibited by even low concentrations of sulfur dioxide. Addition of sulfur dioxide at this stage is also useful in preventing browning reactions.

In Europe the grapes in long-established wine-growing areas are generally so well covered with *Saccharomyces* cells that satisfactory fermentations take place, after an initial light sulfiting of the musts. In United States practice a pure culture of an appropriate *Saccharomyces* is added after the sulfiting, usually as a water-based slurry of commercially available, dehydrated yeast cells.

Yeasts are an important part of the flora of the grape "bloom," but they are not its only members. Bacteria and molds may also be present. Two types of bacteria are specially significant to the winemaker. Certain heterofermentative lactic acid bacteria (*Leuconostoc oenos*), if allowed to multiply, bring about a malo-lactic conversion. They adventitiously convert the dicarboxylic acid, malic acid, of grape juice to the monocarboxylic acid, lactic acid, and carbon dioxide, and at the same time, by other reactions, impart certain fragrance and flavor constituents to the wine. Acetic acid bacteria (*Acetobacter* sp.), if allowed to multiply, oxidize the ethanol in the wine to acetic acid, thus turning the wine to vinegar. The growth of the lactic acid bacteria is encouraged (by the use of as little sulfur dioxide in the sulfiting as possible, and by extended contact of the wine with its lees) by winemakers who use grapes that are not completely ripe at harvest and whose wines would be too tart were it not for the diminution of acidity these bacteria are able to mediate. The growth of acetic acid bacteria is discouraged by keeping the wine away from air as much as possible and keeping the barrel surfaces that are exposed to both wine and air (such as around the bung holes), clean and antiseptic by scrubbing them with a weak solution of sulfur dioxide in water. *See* FERMENTATION; YEAST.

Oxidized wines. Although most wines are deliberately kept away from air as much as possible during cellaring, two types of wine specifically depend on oxidations for their characteristic fragrances and flavors. These are the *vins jaunes* (yellow wines) of the Jura in France, and the sherry wines of the Jerez de la Frontera region of Spain. The oxygen-wine interaction in both cases is mediated by oxidative yeasts (*Saccharomyces beticus*) that form a pellicle (the flor) on the surface of the wine, which is kept deliberately in partly filled casks open to the air. The yeasts oxidize some of the ethanol to acetaldehyde, some of which may form acetals while some also polymerizes to yield other aldehydes, helping to produce the characteristic fragrance of the products.

Although the *vin jaune* of the Jura is vintage-dated, the wines of Jerez are not, these latter reaching the public with a constancy of character that comes about as the result of the wine being of constant average age when released. This condition is achieved by the *solera* system, that is, the uniform fractional blending of younger wines with old. Most so-called sherry wines of the United States are made in an entirely different way: they obtain their characteristic caramellike bouquets and tastes by being heated at 112–150°F (44.5–65.5°C) for several months in the absence of air. Therefore, they would more accurately be known as American "Baked Wine," rather than sherry, since the heating process they undergo is known as baking. Their method of production thus resembles that of the wines of Madeira, which are heated in the same manner.

Sparkling wines. Sparkling wines have, as type species, the effervescent wine of Champagne, in France. The wines from this region are made by allowing a secondary fermentation of dry white still wine (a wine made from only certain grape varieties,

the Pinot noir, Pinot meunier, and the Chardonnay, all not fully ripe when picked) to which sugar and a flocculating yeast have been added. The mixture is put into thick-walled bottles strong enough to withstand more than the internal pressure of the 3–4 atm (300–400 kilopascals) at about 50°F (10°C) in the final wine. Subsequent to its secondary fermentation, by French law Champagne remains in contact with the yeast sediment in the bottle for at least 1 year, before that sediment is moved to the neck of the bottle, there to be frozen when the bottle is prepared for release. At that time the bottle is opened, the frozen sediment removed, and the wine dosed with a solution of sugar in brandy. Champagnes dosed with sugar liqueur are labeled, in order of increasing sweetness, *brut* (very dry), *sec* (dry), *extra-sec* (extra-dry), *demi-doux* (half-sweet), or *doux* (sweet).

Because the clarification of sparkling wine by the Champagne process is labor-intensive, a less expensive transfer process has been introduced into sparkling wine production in areas other than Champagne. In this method the secondary fermentation occurs in small bottles, but at the time the product is being prepared for market the bottles are emptied, under counterpressure, into a collecting vessel, the sediment is removed by filtration, and the clarified wine is dosed and transferred to new bottles. Such wine may be labeled "bottled fermented."

Sparkling wines are also produced by the so-called Charmat or bulk process, in which secondary fermentation takes place in sealed, large (1000–3000 gal or 3800–11,400 liters) glass-lined or stainless steel pressure tanks. The final product must be separated from the yeast soon after the secondary fermentation is completed, to prevent formation of hydrogen sulfide in quantities sufficient to foul the wine. It thus cannot have the autolyzed yeast character that constitutes part of the distinctive flavor and fragrance of Champagne.

Carbonated sparkling wines are made by charging brilliantly clear still wines with carbon dioxide at 32°F (0°C). They cannot usually be aged, however, because they often throw sediments in the bottle, which are stirred up by the escaping gas as soon as the bottle is opened, creating an unsightly product.

Wine quality. The most important factors influencing the quality of wine in a general way are the grape variety and the particular expression of it that is allowed by the environmental conditions of the growing region. While the technical skill of the winemaker is critical, production of distinctive and quality products requires distinctive and quality raw material.

The most esteemed table wine cultivars in the western world are the low-yield but distinctive noble varieties: Chardonnay and White Riesling (each yielding white wine), and Cabernet-Sauvignon and Pinot noir (each yielding red wine). In terms of quality the most suitable regions for growing these varieties are cool regions whose temperature summations lie between 1800 and 2500 degree-days (the degree-day summation is the summation of the difference between the average daily temperature, in °F, for every day of the growing season, and 50°F). Examples of such regions are the famous wine-growing areas of Bordeaux, Burgundy, the Mosel, and the Rheingau in Europe, and the southern portion of the Napa Valley and the peaks of the Santa Cruz Mountains in California. In areas of lower heat summation the grapes will not fully ripen, and in regions of higher heat summation the mature fruit lacks acidity and other needed characteristics. Other environmental factors that influence grape characteristics are vine exposure, soil drainage, and usually to a lesser extent the chemical composition of the vineyard soil.

Edward J. Wawszkiewicz

Bibliography. M. A. Amerine (ed.), *Wine Production and Technology in the United States*, 1981; M. A. Amerine and M. A. Joslyn, *Table Wines: The Technology of Their Production*, 2d ed., 1970; S. F. Anderson and D. Anderson, *Winemaking*, 1989; R. S. Jackson, *Wine Science: Principles and Applications*, 2d ed., 2000; C. S. Ough and M. A. Amerine, *Methods Analysis of Musts and Wines*, 2d ed., 1988.

Wing

The planar surface of a heavier-than-air body whose primary purpose is to produce lift during flight. A wing is composed of a series of airfoil-shaped components joined together to form one integral structure. On an airplane, the wing also provides storage space for the fuel and undercarriage, and is often used as the structural anchor point for the propulsion systems. The wing also includes important flight-control surfaces, such as the aileron, flaps, slats, and spoilers. *See* AERODYNAMIC FORCE; AIRCRAFT PROPULSION; AIRFOIL.

Wing planform definitions. Several key geometric parameters are used to define the wing planform (**Fig. 1**). The chord of a wing section (c) is defined as the straight-line distance between the leading and trailing edges. Other important parameters are the

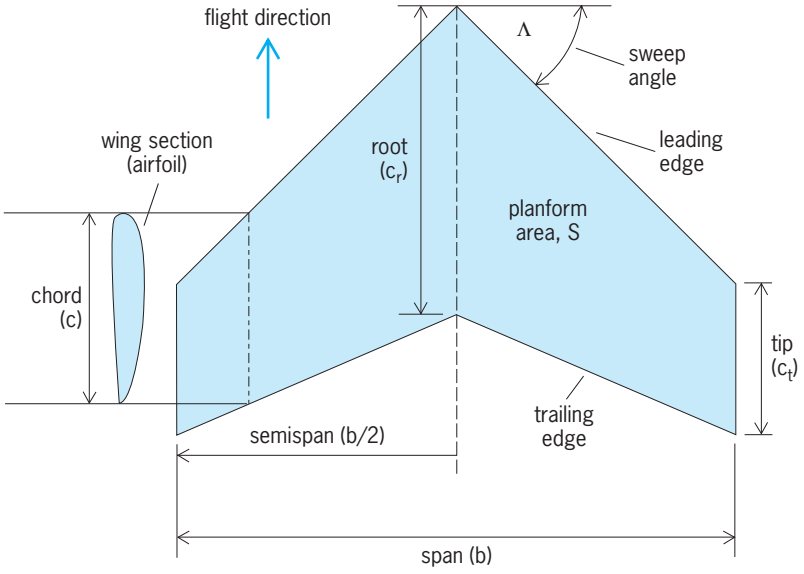


Fig. 1. Wing planform definitions.

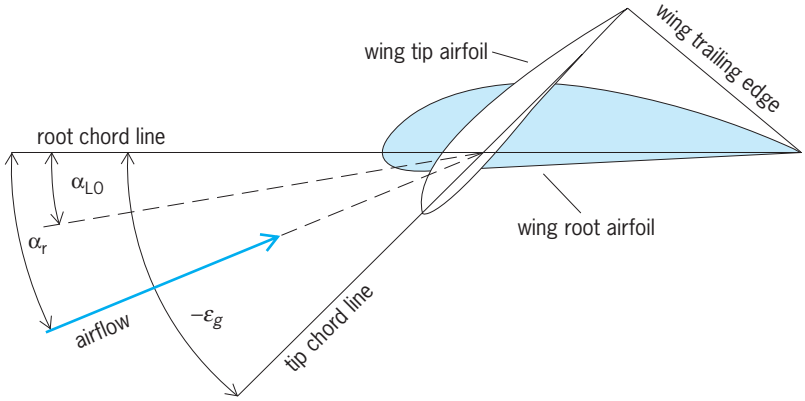


Fig. 2. Wing twist definitions.

leading-edge sweep angle (Λ), the root chord (c_r), the tip chord (c_t), the span (b), the planform area (S), the taper ratio (c_t/c_r), the mean geometric chord ($c_g = S/b$), and the aspect ratio ($A_R = b/c_g = b^2/S$).

Wing twist definitions. The geometric twist at the wing tip (ε_g) is defined as the angle between the root chord and the tip chord (**Fig. 2**). All angles are normally measured in degrees. Twist is defined as positive when the tip is rotated nose up relative to the root chord, and is referred to as wash-in. If the tip is rotated nose down the twist is negative and is called wash-out. As explained later, geometric twist is used to control the air flow over the wing and is, thus, an important wing parameter. The wing angle of attack (α_r) is defined as the angle between the root chord line and the oncoming airflow, known as the free-stream velocity (V_∞). A wing will have a particular angle of attack at which no net lift is produced, and this is known as the wing zero-lift angle (α_{L0}).

Subsonic distribution of lift. It is important to calculate the variation of lift across the wing span, from root to tip, since this is the first step in working out the shear force and bending moment at the wing root which will determine whether the wing structure will fail or not during flight. The shape of the spanwise lift distribution will be influenced by the speed of the air approaching the wing (V_∞) relative to the speed of sound in air (a) which is measured by the Mach number and defined as $M_\infty = V_\infty/a$. If $M_\infty < 1.0$, the flow is classified as subsonic, while, if $M_\infty > 1.0$, the flow is supersonic. The speed of sound decreases slowly with altitude, and has a value at sea level of 340.3 m/s (1116.45 ft/s or 661.5 knots). *See* MACH NUMBER; SOUND.

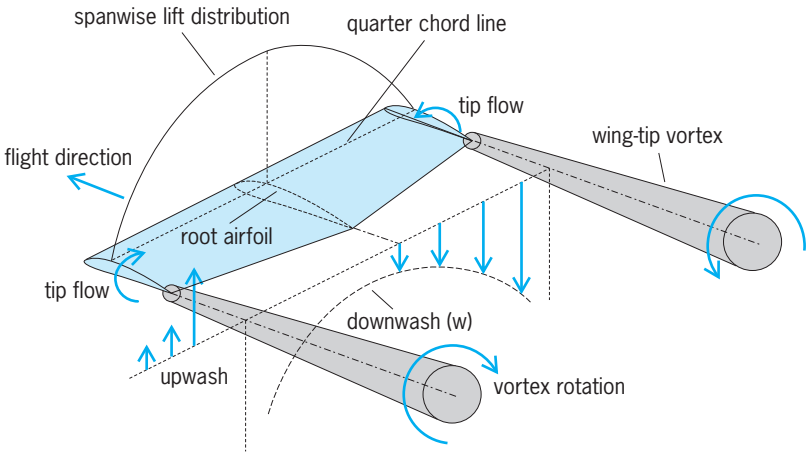


Fig. 3. Subsonic wing-tip vortices.

During subsonic flight, an airfoil produces lift by forcing the air to follow a path over the upper surface which has a larger amount of curvature than that over the lower surface. This flow field causes the air to have a lower static pressure on the upper surface relative to the flow underneath the wing. At each wing tip the lower surface high-pressure air is free to move onto the upper surface, and, as the wing moves forward, two corkscrew flow structures are trailed behind the airplane (**Fig. 3**). These structures are referred to as the wing-tip vortices, and on a humid day they often become visible as water condenses out in the low-pressure vortex core to form two vapor trails behind the airplane. The wing-tip vortices induce a downward flow, known as the downwash (w, measured in the same units as the free-stream velocity), not only behind the wing, but also ahead of the wing. The greater downwash at the wing tip causes the tip airfoil section to experience a much lower angle of attack than the root airfoil, which results in the lift being zero at the tips and a maximum at the wing root (Fig. 3). In the period 1912–1918, Ludwig Prandtl and his colleagues in Germany developed a mathematical model of subsonic flow, known as lifting-line theory, which allows the spanwise variation of lift to be calculated for a given wing geometry, the first step in obtaining the important wing root shear force and bending moment. *See* BERNOULLI'S THEOREM; VORTEX; WAKE FLOW.

Downwash also gives rise to a drag component known as the induced drag (or drag due to lift), which increases as the wing aspect ratio decreases. At low angles of attack, wings which have a large aspect ratio (such as those of gliders and the Lockheed U2) produce more lift for a lower induced drag than wings of low aspect ratio. Thus, high-aspect-ratio wings have a higher lift-to-drag ratio and are often referred to as being aerodynamically efficient. *See* SUBSONIC FLIGHT.

Wing sweep. As the air flows over the wing, changes in static air pressure are transmitted through the air at the speed of sound. If the airspeed anywhere over the wing approaches the speed of sound, these pressure signals cannot travel farther upstream than this point and add together to form a shock wave at this location. If the pressure signal is small, a very weak shock wave forms, which is known as a Mach wave. When the air passes through a shock wave its velocity decreases and the static pressure increases. One undesirable outcome of the formation of shock waves over the wing is a large increase in drag on the airplane. *See* AERODYNAMIC WAVE DRAG; SHOCK WAVE.

On many aircraft, the leading edge of the wing is swept back by an angle Λ degrees (Fig. 1). The effect of sweep relies on the fact that it is only the airspeed normal to the leading edge which controls the

variation of air velocity over the upper and lower surfaces of the wing. Thus, even though the free-stream Mach number (M_∞) may be equal to unity, the airspeed normal to the leading edge ($M_\infty \cos \Lambda$) may be small enough to avoid increased drag due to shock-wave formation over the wing. Typical values of sweepback for modern transport aircraft range from 25 to 37°. *See* TRANSONIC FLIGHT.

Supersonic distribution of lift. When a wing flies at supersonic speeds ($M_\infty > 1.0$), any source of pressure disturbance can only influence the flow within a downstream conical volume called the Mach cone which extends from the source and whose curved surface is defined by Mach waves (very weak shock waves). Mach waves slope rearwards at an angle of sin $\sigma_M = (1/M_\infty)$ to the horizontal (**Fig. 4**). One major difference between subsonic and supersonic wing flow is that in supersonic flow the downwash of the wing tip vortices is only felt within the Mach cone and not over the entire wing span or fuselage (Fig. 4). Identifying the regions which are, or are not, influenced by downwash is an important first step in calculating the spanwise distribution of lift which controls the structural loads on the wing during flight. *See* SUPERSONIC FLIGHT.

Wing stall. A wing is said to be stalled when the lift starts to decrease for an increase in angle of attack. In general, this happens when the flow ceases to follow the exact contour of the airfoil geometry, which is known as flow separation. The flow over airfoil A in **Fig. 5** has flow separation at the trailing edge. As the wing angle of attack is increased, the separation spreads over a region which is often referred to as a stall cell. Roll control of the airplane is achieved using control surfaces known as ailerons. These surfaces increase or decrease the camber of the wing section (Fig. 5) which, respectively, increases or decreases the lift. If the lift is increased on the right wing and decreased on the left, the airplane will bank (roll around the fuselage centerline) and begin to fly in a curved path. A wing is normally tapered to reduce the wing root structural loads due to weight, and rear swept to delay the formation of shock waves over the wing upper surface during cruise. However, both taper and sweep increase the likelihood of the stall cell first forming over the ailerons. By introducing negative geometric wing twist, the stall cell is forced inboard toward the wing root, essentially by allowing the wing tips to fly at a lower angle of attack than the root. This maintains smooth airflow over the ailerons and allows the pilot to maintain roll control of the airplane at all times. A similar effect to geometric twist can be achieved by carefully varying the wing airfoil section between root and tip; this is called aerodynamic twist. *See* AILERON.

Wing loading. The total lift produced by a wing is given by $L = \frac{1}{2}\rho_\infty V_\infty^2 S C_L$ where C_L is the wing lift coefficient and is dimensionless. When the air density ρ_∞ is measured in kg/m^3, V_∞ in m/s, and S in m^2, the lift has the units of newton, N. For a given value of free-stream velocity (V_∞) and wing planform area (S), the wing lift coefficient increases with angle of attack (α_r) up until the stall point; this data may be obtained from wind-tunnel tests of a scaled-down model of the wing. It is important to note that, for a given angle of attack, the wing lift is controlled by the square of the flight speed and the wing area.

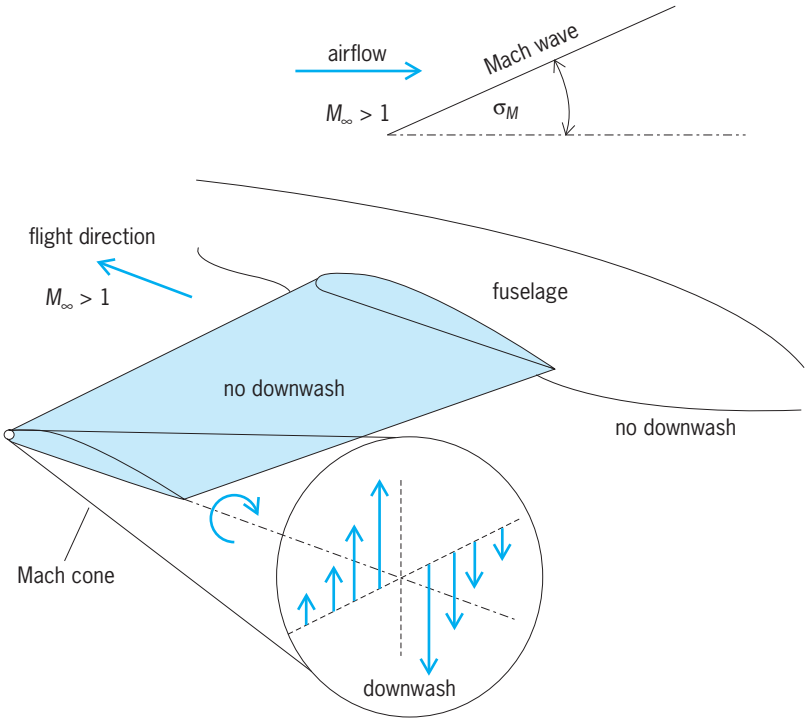


Fig. 4. Supersonic downwash. Inset shows geometry of Mach wave.

Wing loading is defined as the airplane mass (W) divided by the wing planform area (S). During takeoff, where the airspeed is low, a large wing area is required to produce sufficient lift to get the airplane airborne, and this results in a low wing loading. However, drag on the wing increases with planform area. A large drag at cruise would be undesirable, especially since, at the higher airspeed, a

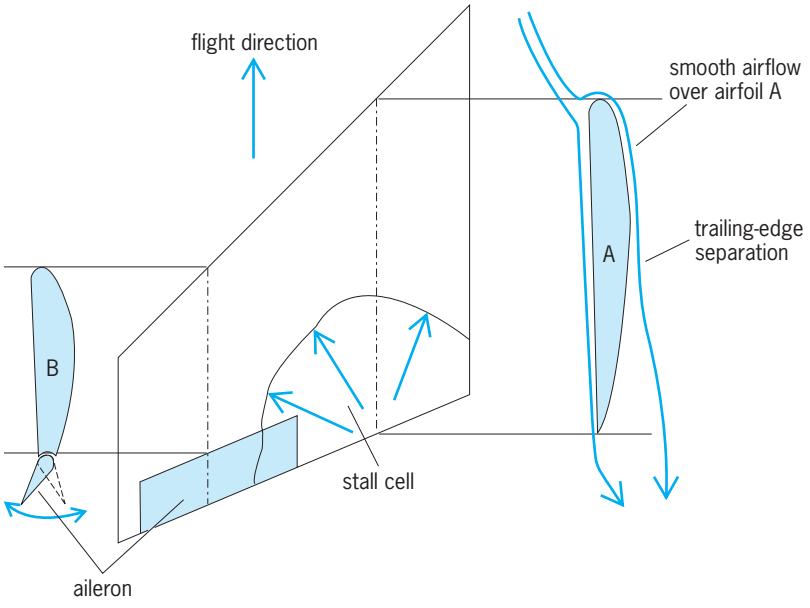


Fig. 5. Stall cell formation.

Typical wing loadings for birds and airplanes

Example	Wing loading, kg/m^2*
Swallow	1.5
Hummingbird	2.4
Golden eagle	7.1
Canada goose	20
General aviation aircraft	Values around 100
Commercial passenger aircraft	100 to 150
Combat aircraft	360 to 580
F-16 fighter aircraft	361

*1 kg/m^2 = 0.2049 lb/ft^2.

smaller wing area could produce the required lift with a higher wing loading, and aircraft with a high wing loading have a greater maximum airspeed than those with lower wing loadings. A compromise is made by increasing the wing area during takeoff and landing through the use of high-lift devices. Typical wing loadings of birds and airplanes are given in the **table**.

High-lift devices. Wing high-lift devices, specifically referred to as slats and flaps, are deployed during take-off and landing when high lift is required at slow flight speeds (**Fig. 6**). However, they are not used during cruise (Fig. 6*a*), since they greatly increase the drag on the airplane. When deployed, these control surfaces increase both the camber and chord of the cruise airfoil section (Fig. 6*b*). Increasing the camber increases the curvature of the air as it flows over the airfoil, which further decreases the upper-surface static air pressure (over the cruise configuration) and, thus, increases wing lift. The increased chord increases the wing planform area (S), which also increases wing lift. After touchdown, spoilers are deployed, which increase the drag on the airplane and quickly reduce the lift, thus applying the weight of the airplane on the braking wheels. Spoilers can also be used individually during high-speed flight to reduce lift on one wing only and thus roll the aircraft, and, when used together, to act as an airbrake to slow the aircraft quickly. *See* AIRPLANE; WING STRUCTURE.

Andrew J. Niven

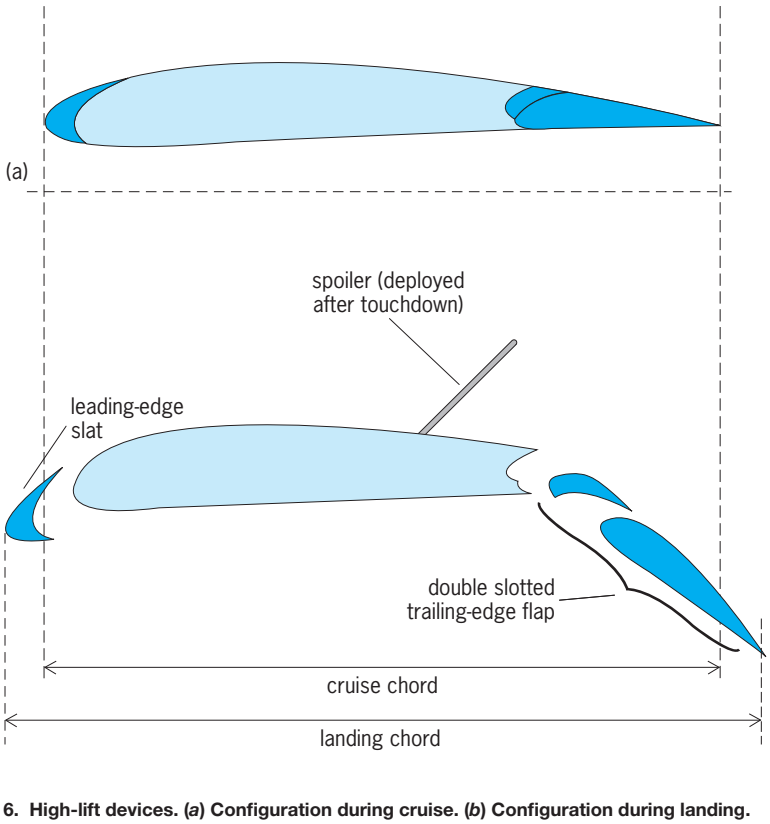


Fig. 6. High-lift devices. (*a*) Configuration during cruise. (*b*) Configuration during landing.

Bibliography. I. H. Abbott and A. E. von Doenhoff, *Theory of Wing Sections*, Dover, 1949; H. Ashley and M. Landahl, *Aerodynamics of Wings and Bodies*, Dover, 1985; J. D. Anderson, Jr., *Introduction to Flight*, 4th ed., McGraw-Hill, 2000; M. Davies (ed.), *The Standard Handbook for Aeronautical and Astronautical Engineers*, McGraw-Hill, 2002; H. Schlichting and E. Truckenbrodt, *Aerodynamics of the Airplane*, McGraw-Hill, 1979; D. Stinton, *The Design of the Aeroplane*, 2d ed., Blackwell Science, 2001.

Wing structure

In an aircraft, the combination of outside fairing panels that provide the aerodynamic lifting surfaces and the inside supporting members that transmit the lifting force to the fuselage. The structure of a wing is an integration of the environment external to the vehicle wing, the aerodynamic shape of the wing, and the proposed use of the vehicle. The interaction of these three aspects of design leads to the selection of material, to the general structural layout, and, finally, to the detailed choice of structural shapes, material thickness, joints, and attachments. The result is a structural framework covered with a metal skin that also contributes to the load-carrying function. *See* SUPERCRITICAL WING; WING.

Structural materials. Wing structure has evolved from the early use of wood, doped canvas, and wire. The first general change was the replacement of wood with metal frameworks; wood and canvas structure is still found, however, in lightweight personal aircraft. Doped canvas was replaced by light metal skins, which served only as fairing in the beginning but which were later designed to provide a portion of the structural strength. Today, aluminum alloy outer skins are prime structural elements on all commercial transports and on the great majority of military craft.

Magnesium, steel, and titanium are also used in internal primary structure and in local skin areas. For materials used in the wing structure of research and other exotic aircraft *see* ATMOSPHERIC ENTRY.

Design of a wing structure begins logically with a derivation of loads on the wing, both flight and net, as they are affected by particular design specifications; then the design proceeds through a choice of material to the final phases, when the configurations of major structural units such as the prime wing box

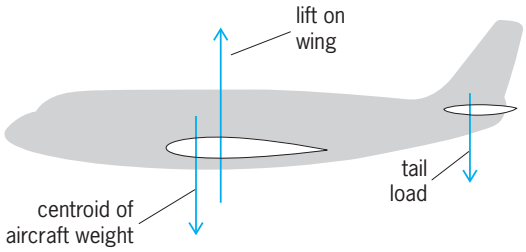


Fig. 1. Schematic diagram of major air loads.

and the leading- and trailing-edge subassemblies are determined.

Total flight load. Flow of air past the wing's airfoil shape creates a lifting force. Total lifting force L, in pounds (newtons), on an entire wing is given by the equation below, where ρ is air density in slugs per cubic foot (kilograms per cubic meter), V is velocity

$$L = \frac{\rho V^2 C_L S}{2}$$

in feet per second (meters per second), S is wing area in square feet (square meters), and C_L is lift coefficient, which is a nondimensional coefficient derived from wind tunnel tests. *See* AERODYNAMIC FORCE; AIRFOIL.

On a particular craft, the lifting force required of a wing depends on the craft's balance. For stability, the aircraft is designed to maintain the centroid of any permissible loading condition forward of the center of lift on the total wing (**Fig. 1**).

The total load on the wing in normal cruising flight is the weight of the aircraft plus the balancing tail load. The total maximum anticipated load in service is much greater than this because the aircraft will encounter atmospheric gusts and because it is required to maneuver. The designed breaking strength of the structure is normally set at 1.5 times the maximum anticipated loads, the 1.5 multiplier being a factor of safety. The resulting load is called the ultimate design load.

Maneuver loads. An aircraft must be capable of climbing, descending, turning, or, if it is a fighter, of executing rather violent acrobatics. Any such changes from a statically balanced condition produce load increments. These loads are generated when the pilot induces a sudden change in the angle of attack of the aircraft by manipulating the engine throttles and controls. An abrupt change in speed or angle of attack alters the lift coefficient or velocity and produces a change in total lift. The permissible degree of maneuverability of any specific model is based on its intended use. The **table** presents general information on ranges of maximum anticipated load for various types of craft. Down increments of maneuver load are limited to two times the weight for transports and to somewhat higher values for the others. The ultimate design load is 1.5 times the value shown. *See* FLIGHT CONTROLS.

Maximum anticipated upward flight loads

Type of craft	Required lift force on craft
Transport	2.5 × weight
Bomber	2.5–4 × weight
Fighter	6–9 × weight
Special research	Up to 12 × weight

The multipliers in the table, which are called load factors, reflect the intended usage of the craft. Experience has shown these to be adequate, but the pilot must be aware of the maneuver limitations of the craft. It is possible to exceed these loads in flight. **Figure 2** shows a simplified presentation of load factor and factor of safety applied to total external wing load. Tail load is assumed zero for simplicity of presentation. The presentation shows that the external air load equals the aircraft weight during cruise. In a 2.5-g maneuver, maximum anticipated air load is 2.5 times the aircraft weight, the 2.5 being a design factor. On this basis, ultimate design load is 1.5 × 2.5, or 3.75, times the aircraft weight, the 1.5 being the safety factor.

Gust loads. A second significant load produced on any craft's wings is an atmospheric gust. A high-velocity vertical gust produces an abrupt change in the direction of airflow over the wing. An upward gust produces an upward incremental force, and a down gust produces a downward incremental gust. The U.S. Civil Aeronautics Board (CAB) specifies the multiplying gust load factor.

For commercial aircraft, the normal gust velocity in feet per second (1 ft/s = 0.3 m/s) is specified as $30K$ at cruise speed, $15K$ at dive speed, and $40K$ for the speed recommended for flying under more turbulent conditions (less than cruise speed), where K is a coefficient dependent on wing loading, which is the ratio of aircraft weight to wing area. The recommended speed during turbulent conditions is determined from the aerodynamics of each aircraft type. The value K is about 1.20 for many typical current designs. The factors must be derived for both upward and downward gusts, which produce loads on

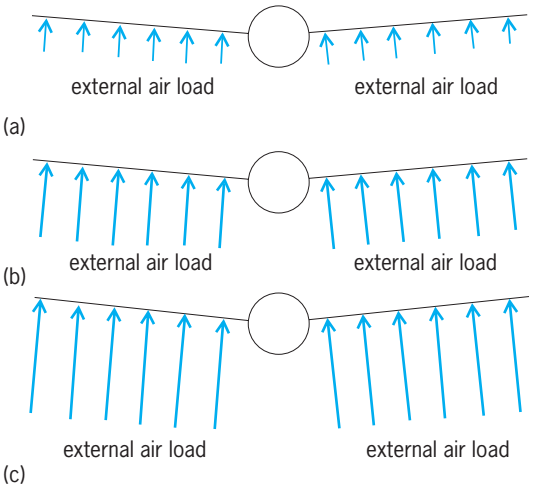


Fig. 2. Load factor and safety factor applied to total external air load. (*a*) External air load during cruise. (*b*) Maximum anticipated air load in 2. 5-*g* maneuver. (*c*) Ultimate design load with 1.5 factor of safety.

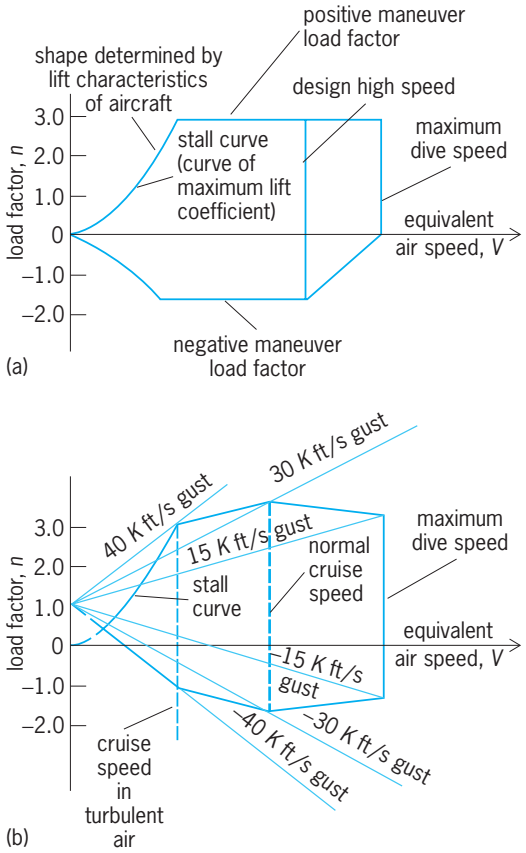


Fig. 3. ***V-n*** **diagrams of maximum anticipated flight loads. (*a*) Maneuver envelopes. (*b*) Gust envelope. 1 ft/s = 0.3 m/s.**

the wing comparable to the up and down maneuver loads. The most critical among these loads must be determined for each element of the wing structure.

V-n diagram. The design speeds and load factors for any design are summarized in a graph called the $V-n$ diagram (**Fig. 3**). The curve is derived from the specified requirements for the particular use of the vehicle. The curve applies to the craft as a whole. Balance considerations depicted in Fig. 1 permit the derivation of the total flight loads.

Landing and takeoff loads. Landing and takeoff loads are specified as part of the CAB air regulation for commercial vehicles and in connection with ground loads for military vehicles. Commercial aircraft must be capable of anticipated descent velocities of 10 ft/s (3 m/s) at design landing weight and 6 ft/s (1.8 m/s) at design takeoff weight. These descent velocities must be considered at different landing attitudes of the craft. Anticipated weight and centroids of design conditions are established by considering the vehicle usage. In addition, one-wheel landings, lateral drift, braking, turning, and pivoting are evaluated and defined in terms of loads. The dynamic effects of spinning the landing gear wheels at touchdown and the subsequent springback of the gear and its attaching structure are critical conditions that are also evaluated. In general, the specified loadings and state-of-the-art development of landing gear shock-absorbing devices produce actual vertical loads on each main gear approximately equal to the aircraft weight. Maximum aft loads on the main gear are of the same order of magnitude, and side loads are 60–80% of this. This load envelope at landing usually determines the design of structure in the wing to which the main gear attaches. Because of the large twisting forces induced by the drag loads applied at ground level, which must be carried up and into the wing, this envelope is also critical for significant portions of the inboard wing structure. Ten or so landing conditions affect specific portions of the wing. The ultimate reaction of these loads is, of course, the weight and rotational inertia of the fuselage. The gear is expected to withstand actual anticipated loads without permanent distortion and to withstand ultimate design loads without structural failure. *See* LANDING GEAR.

Net wing loads. The maneuvers and gusts acting on the craft also generate accelerations and inertia effects. Therefore, the net loads on the wing, which the structure must sustain, are the net sum of the previously derived loads and the counteracting inertial effects. A simple example of a flight condition is presented in **Fig. 4** for a hypothetical 100,000-lb (45,000-kg) aircraft in which the tail load is assumed to be negligible. In essence, the wing is first assumed to be isolated from the rest of the craft. Then the air loads and inertia loads are applied, and the net of these is balanced with the remainder of the craft and applied to the wing through the wing-to-fuselage connections.

From 15 to 25 different flight conditions are normally derived to satisfy possible critical conditions. These flight conditions encompass variations of

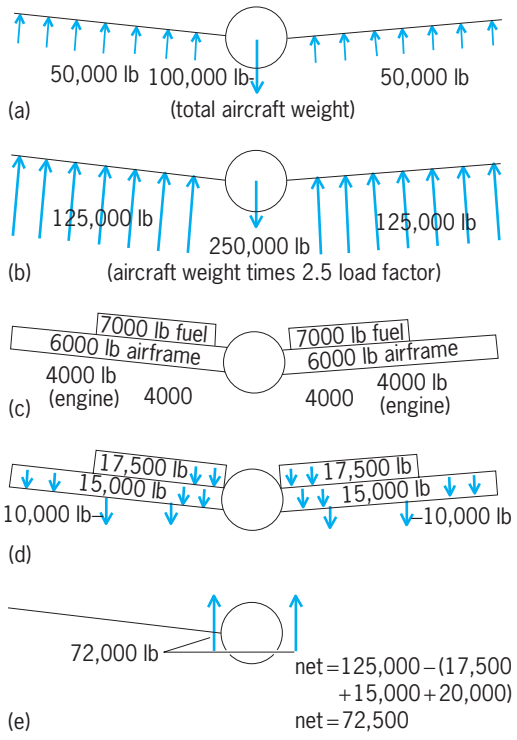


Fig. 4. Derivation of net loads. (*a*) External air load during normal cruise. (*b*) Total wing load during 2.5-*g* maneuver. (*c*) Wing dead weight at 1 *g*. (*d*) Wing weight times 2.5 load factor. (*e*) Net shear load to fuselage. 1 lb = 0.45 kg.

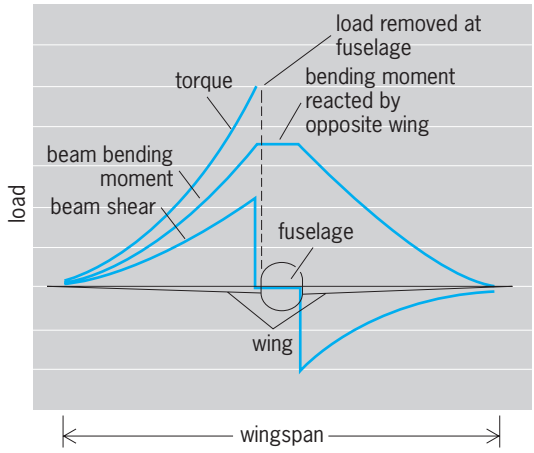


Fig. 5. Shear, bending moment, and torque curves.

aircraft gross weight and centroid location, fuel loadings, upward and downward maneuvers and gusts, special control surface and landing flap conditions, and any unusual dynamic loadings, such as dynamic interactions of unsteady airloads and structural stiffness. Dynamic loads are also affected by concentrated weights such as engines, tip tanks, special pods, and gun or rocket reactions. *See* AEROELASTICITY.

Wing structure is examined at a number of wing cross sections in order to taper the weight of the material from the tip to the inner section of the wing. This is done by summing the net loads from the tip of the wing to the section in question. This calculation yields the vertical and drag shear forces, the bending forces produced by them as they are carried in toward the fuselage, and the twisting or torsion forces produced by an eccentricity of the vertical and drag forces about the twisting center of the structure. These forces are presented as beam (vertical) and chord (drag) shear curves, beam and chord bending moments, and torque (**Fig. 5**).

Choice of material. The derivation of the net loads permits a quantitative consideration of the general structural framework. Design of the structure, however, is inseparable from the choice of material. The thickness and shape of the individual pieces also affect strength in lightweight, efficient structures because of compression stability modes.

High static strength and light weight—coupled with reasonable rigidity characteristics, good corrosion resistance, fatigue endurance, cost and formability—are prime considerations in choosing airframe material. Highly loaded, efficient airframes built since 1930 have found these qualities in aluminum. Alumium has withstood the technical competition of steel, titanium, magnesium, wood, and reinforced plastics, although complete designs have been manufactured from each of these competitive materials to try them with complete realism. At present, composite materials are being introduced into both military and civilian aircraft at an increasing rate. *See* COMPOSITE MATERIAL.

Aluminum has been somewhat supplanted by the requirements for high-supersonic and hypersonic speeds in the atmosphere (the thermal thicket) and by the demands of space technology (the atmospheric reentry problem). But even in the case of aerospace vehicles with these problems, designs incorporating protective shields and cooling of the structure have been considered to permit the use of aluminum with its favorable static strength, corrosion resistance, fatigue strength, cost, and formability, at so light a weight. *See* AEROTHERMODYNAMICS; HYPERSONIC FLIGHT.

It does appear, however, that the demands for speed interacting with the physical environment are causing aluminum to be replaced by titanium and other alloys to some extent. A material's creep characteristics and strength at high temperature are now important added parameters.

Authoritative data on strength properties of aircraft metals are available. **Figures 6–9** present in

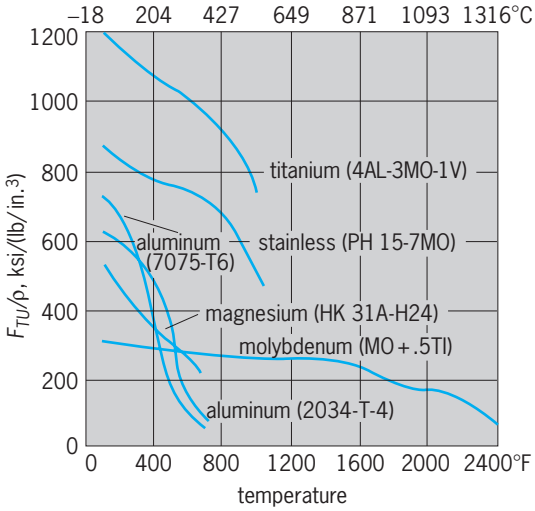


Fig. 6. Graph showing ultimate tensile strength F_{TU} divided by density ρ for various materials. 1 ksi/(lb/in.^{3}) = 0.25 MPa/(g/cm^3).

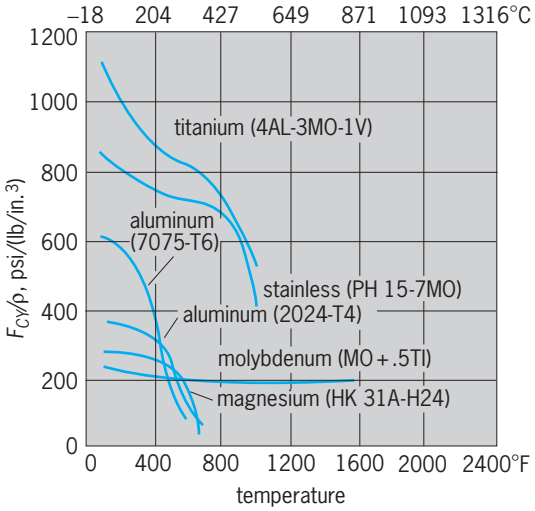


Fig. 7. Graph showing compressive yield strength F_{CY} divided by density ρ for various materials. 1 psi/(lb/in.^{3}) = 0.25 kPa/(g/cm^3).

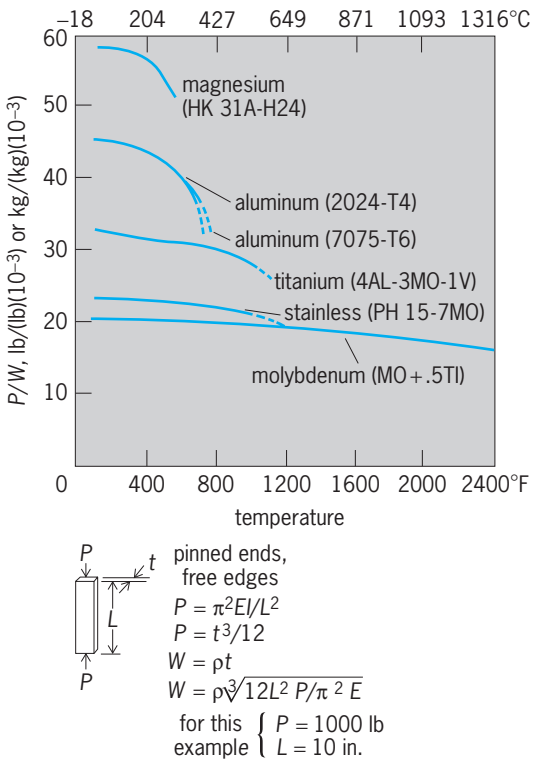


Fig. 8. Critical compression load *P* divided by weight *W* in elastic region (Euler column) for various materials. 1000 lb = 454 kg. 10 in. = 25.4 cm.

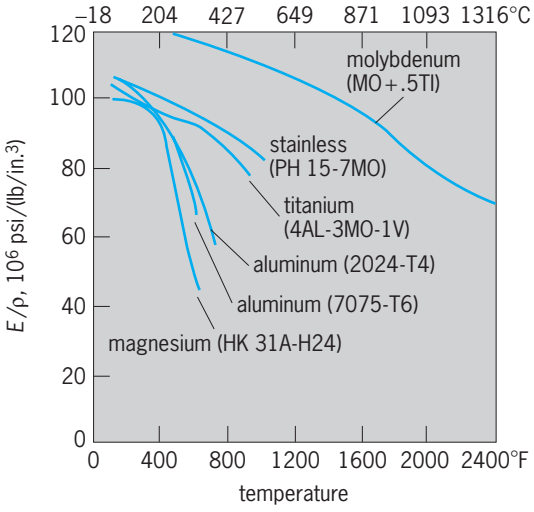


Fig. 9. Modulus of elasticity *E* divided by density ρ for various materials. 106 psi/(lb/in.^{3}) = 0.25 GPa/(g/cm^3).

graphical form comparisons of strength and stiffness with temperature. These curves indicate the superiority of aluminum at room temperature, and the disparity among materials as the temperature increases. For example, the ultimate tensile strength F_{TU} per unit weight is useful for application where tension loads predominate, such as on the lower surface of the wing (Fig. 6). The favorable tension ratios of titanium and molybdenum are negated by cost or formability, as well as by consideration of all properties shown in Figs. 7, 8, and 9. The compressive yield strength (F_{CY}) per unit weight is useful where compressive stresses and compression stability considerations are prime factors, such as on the upper surface of the wing (Figs. 7 and 8). The plot of Young's modulus of elasticity per unit weight is a useful parameter for stiffness considerations (Fig. 9). *See* STRESS AND STRAIN; YOUNG'S MODULUS.

Hypersonic flight in the atmosphere and the reentry problem emphasize characteristics such as strength at high temperature, thermal conductivity, specific heat, coefficient of expansion, melting point, creep, and oxidation temperature, as well as the properties already mentioned. Data on the thermophysical properties for many of the solid materials that may be used in the future are being measured.

Prime structural framework. An airframe wing is essentially two cantilever beams joined together. Each wing tip is the free end of the cantilever, and the centerline of the vehicle represents the plane where the two fixed ends of the cantilevers are joined (**Fig. 10**). The prime load-carrying portion of these cantilevers is a box beam made up usually of two or more vertical webs, plus a major portion of the upper and lower skins of the wing, which serve as chords of the beam. This box section also provides torsional strength and rigidity (**Fig. 11**). Normally the prime box is designed to carry all the primary structural loads; these include all beam shears and bending moments, all drag shears and bending moments, and the torsional or twisting loads. *See* CANTILEVER.

Leading and trailing edge portions of the wing, forward and aft of the prime box respectively, help to provide the airfoil shape required. These portions are designed to minimize their participation in the major load-carrying function. Where participation is forced by the detail design, the fasteners and

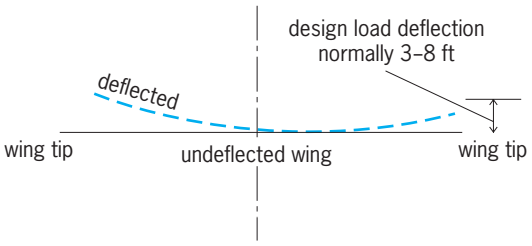


Fig. 10. Wing tips deflect upward in normal flight. 3–8 ft = 0.9–2.4 m.

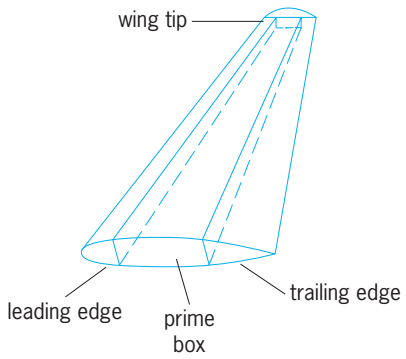


Fig. 11. Prime box of wing half-span.

materials reflect this, but normally the prime box strength is not reduced. An overlapping conservative assumption thus results.

Beam shear material. The vertical webs of box beams are the prime load path for the beam or vertical shears. These webs are comparable functionally to the webs of plate girders in highway or railroad bridges. The prime difference is the elastic buckling permitted in the aircraft web and consequently the much lighter material thickness. This buckling is clearly shown in **Fig. 12**, a photograph of a major aircraft beam under laboratory test where the shears imposed are buckling the web. When the load is removed, the buckles will disappear. This type of beam is called a partial tension field beam. *See* PLATE GIRDER.

The web material of these beams is designed by net flight loads in the outer portion of the wing. The beam and torsional or twisting loads will normally combine to produce the loads which design the webs near the front of the wing. This is usually a large angle of attack condition. A low angle of attack condition, or one involving use of the control surfaces in the trailing edge region, normally produces design shears in the rear webs. In the inner portions of the wing, especially inboard of the landing gear, the landing conditions produce shears which exceed the flight loads by significant amounts. These loadings require substantially thicker webs. Aluminum webs of 0.020 in. (0.5 mm) thickness may be anticipated in the outer portions of the wing. These increase toward the wing root, and thicknesses of 0.125–0.188 in. (3.2–4.8 mm) may occur at the inner section.

One of the beams from a twin-engine transport is shown in **Fig. 13.** A portion of the beam at the centerline of the aircraft is of truss construction to permit access for equipment and for inspection and maintenance.

Top of wing box. The top of the prime box is the compression chord of the cantilever beam. It is also the portion of the box where the greatest variations in construction are found among the products of various manufacturers. The desirability of minimum weight and the requirements of the various positive and negative bending conditions have produced numerous configurations (**Fig. 14**).

These cover structures are designed as beam columns or plate structures. They are subjected to axial load by the beam and chord bending moments, coupled with lateral loads from air pressures and, if applicable, fuel tank pressures. The final design is achieved by an iterative process. Approximate material sizes are established for a complete cross section, that is, for top of the wing box, bottom, and connecting corner members. Detail stresses are calculated by methods fully described in texts. Repetitive refinement of material sizes finally results in an efficient top cover based on minimum weight and practical construction methods. Here, too, the gages begin with a minimum of 0.010-in. (0.25-mm) aluminum in honeycomb or 0.020-in. (0.5-mm) in corrugated forms. The plate designs begin with 0.050- to 0.060-in. (1.25- to 1.5-mm) minimum and reach thicknesses of 0.750 in. (19 mm) at inboard sections on some of the largest craft. The basic material thicknesses are determined for the compression loads and are virtually always adequate for load reversals which would put the top of the box in tension. Frequently, elements of the top of the box, including splices, are tested during the development of the craft to ensure adequate strength for the final design. *See* AIRCRAFT TESTING; BEAM COLUMN.

Fig. 12. Partial tension field beam under test.

Fig. 13. Center section of main wing beam of a commercial transport.

Bottom of wing box. The bottom surface of the prime box is the tension chord of the beam. The bottom surface is most frequently a skin, tapering in thickness from a minimum at the wing tip to much heavier gages at the root. Specific gages are on the order of two-thirds to three-fourths as thick as those for the top cover of the wing box because the allowable tensions are always higher than the permissible stresses for compression stability. Fatigue strength is a significant consideration in bottom surface design. Analyses and tests are used to provide the most reliable answers available. Spanwise reinforcing members on the bottom surface bring the compression strength up to requirements for load reversals. The lower surface must normally carry 35–40% of its tension allowable as a compression design load. Framing members around access openings help carry the stresses around such structural discontinuities in the lower wing frame (**Fig. 15**).

Wing ribs. At numerous places within the wing box, bulkhead-type structures called ribs are located.

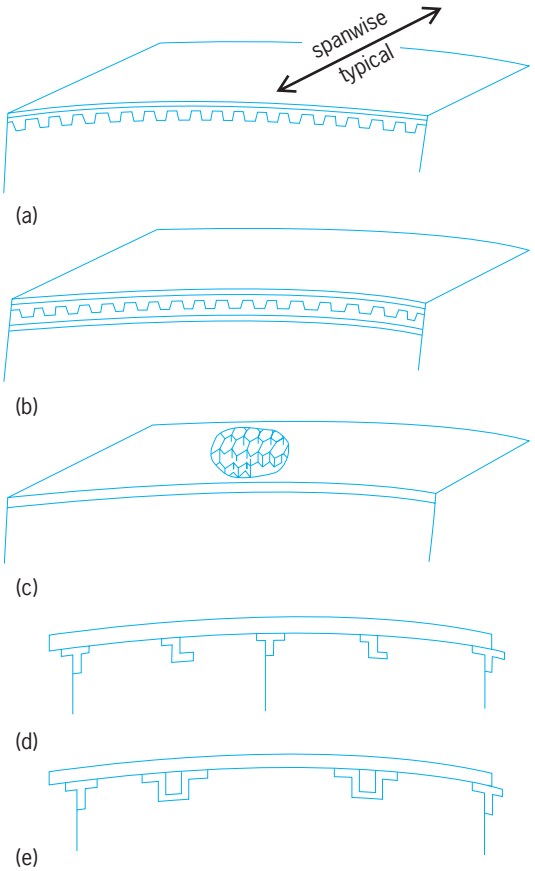


Fig. 14. Types of top cover structure. (*a*) Single skin and corrugation. (*b*) Double skin and corrugation. (*c*) Honeycomb with two skins. (*d*) Thick plate and intermediate stringer. (*e*) Plate with relatively heavy stringers.

These internal structures serve to maintain the rectangular box shape and to cut down the unsupported length of compression cover structures, to separate fuel tanks, and to distribute concentrated loads from guns, bombs, landing gear, or engines into the prime box. They are also located at any wing cross section where major load redistributions occur.

Fig. 15. Lower surface framework of outer wing of a commercial transport.

Wing weights and statistics. The weight of the aircraft divided by the wing area is called the wing load. From 1940 to 1960 wing loadings steadily increased from 35 to 100 lb/ft^2 (171 to 488 kg/m^2) on high-speed aircraft, and they have continued to increase since then.

While wing loading or intensity of loading has been increasing, the thickness of the wing has decreased from approximately 10% of chord (the distance from the leading edge to trailing edge) to 5% of chord. Thus, higher loadings on a thinner beam have increased the challenge to the designer to provide more strength in a shallower space. This has been done at the cost of increasing weights. In 1940, 4 lb/ft^2 (20 kg/m^2) represented a reasonable allowance for average wing weight. In 1960, a high-speed aircraft required about 10 lb/ft^2 (49 kg/m^2). Supersonic speeds demand specific weights well above 10 lb/ft^2.

Leading-edge structure. The leading edge, or most forward portion of the wing, serves an important aerodynamic function in establishing smooth airflow and efficient lifting power for the wing. This area sustains the highest aerodynamic pressures.

Structurally, the leading edge is an appendage, a fairing whose local loads must be supported by the prime wing box. This structure takes the form of cantilever beams or arch-type structures from the leading edge back to the front vertical beam (**Fig. 16**). These leading-edge structures are normally of 0.032- to 0.972-in. (0.8- to 1.8-mm) aluminum, with skin covers in the same thickness range.

Frequently the leading edge of an aircraft incorporates anti-icing provisions. Two prime systems have been used in the past. One type is a rubber bladder that is stretched over the leading edge and is alternately inflated slightly and deflated. This action cracks any ice formation, and the airstream sweeps it away. A second method is the hot leading edge. In this system, air that is heated by special heaters or by the engine exhaust is passed along the leading edge to melt the ice. This heated air normally exhausts at the wing tip.

Numerous aircraft have leading-edge slats. These small airfoil sections are pulled out and forward automatically by air pressures and serve to maintain smooth airflow to a higher angle of attack than the basic large airfoil would sustain. As the angle of attack is decreased, these devices are withdrawn into the basic airfoil by springs (**Fig. 17**). Slats are usually designed in short two-hinge lengths to eliminate wing deflection effects.

Trailing-edge structures. The trailing-edge structure is noteworthy for its various auxiliary devices which assist in aircraft control or in reducing aircraft landing speeds. Various forms of flaps, ailerons, and spoilers, with their hydraulic or electrical control mechanisms, fill the volume of space aft of the rear beam or spar (**Fig. 18**).

Landing flaps. Landing flaps consist of a movable airfoil-shaped structure located aft of the rear beam or spar. They extend about two-thirds of the span of the wing. Their aerodynamic function is to increase

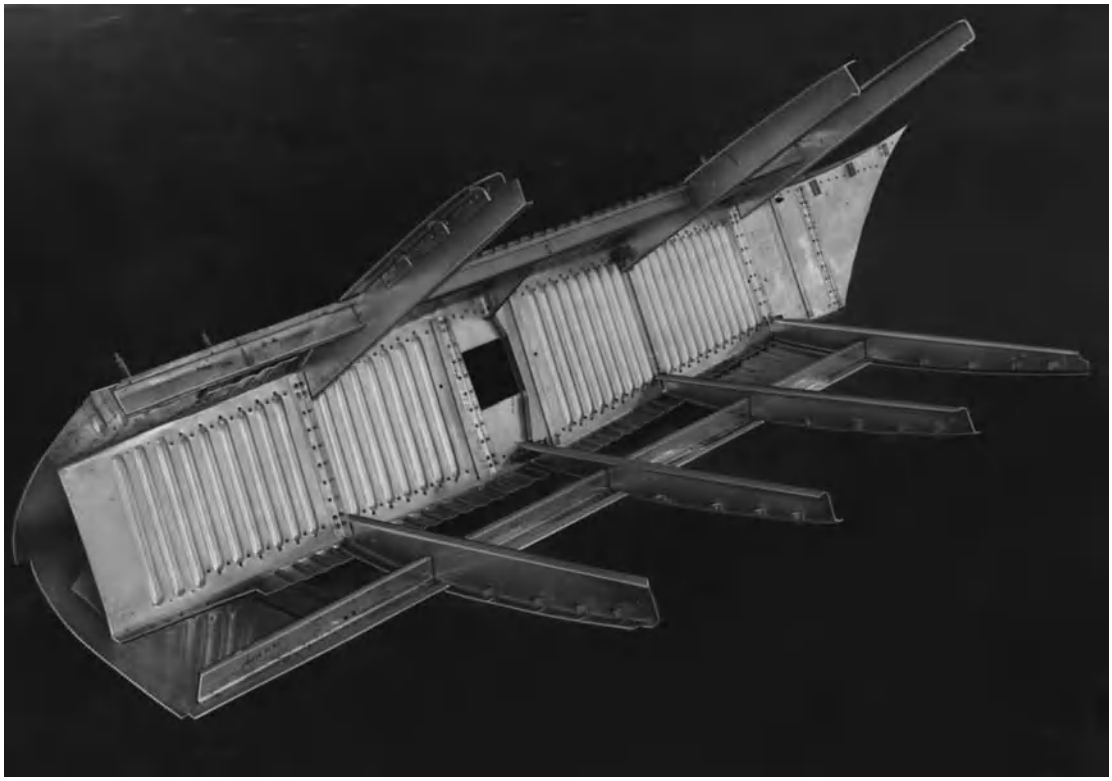


Fig. 16. Section of wing leading edge of a commercial transport.

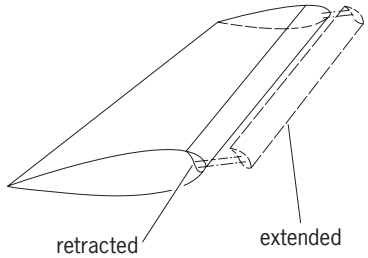


Fig. 17. Leading-edge slat is usually in short two-hinge lengths.

substantially the lift, thereby permitting lower takeoff and landing speeds.

Flaps may be hung from the wing on two, three, or four hinges. Structurally they represent a simple or continuous beam, depending on the number of hinges. Where more than two hinges exist, the wing deflection effects must be considered in addition to local air loads. That is, as the wing deflects it loads the flap in bending it to the shape of the deflected wing. Flaps normally have a single spanwise main beam. Leading-edge ribs and trailing-edge ribs are attached to this main beam, and the airfoil shape is skinned over (**Fig. 19**). Normal gages in these aluminum structures are between 0.020 and 0.040 in. (0.5 and 1.0 mm). The hinge supports are cast or forged aluminum members. The hinges may be plain or roller bearings. *See* ELEVATOR (AIRCRAFT).

Ailerons. Ailerons are located near the tips of the wings in the trailing edge. Their prime aerodynamic function is roll control of the aircraft. Structurally, the aileron is similar to the landing flap. *See* AILERON.

Spoilers. Spoilers are a special form of control surface and perform the aerodynamic function of the aileron, that is, roll control. They are located on the upper surface of the trailing edge at about midsemispan and deflect only upward by pilot control. The cross section of the spoiler is usually rectangular. It is a small box beam consisting of two spanwise beams plus ribs and covering skin.

Building the trailing edge. The various aerodynamic surfaces in the trailing edge region require 5–20 hinges whose locations must be closely controlled. The

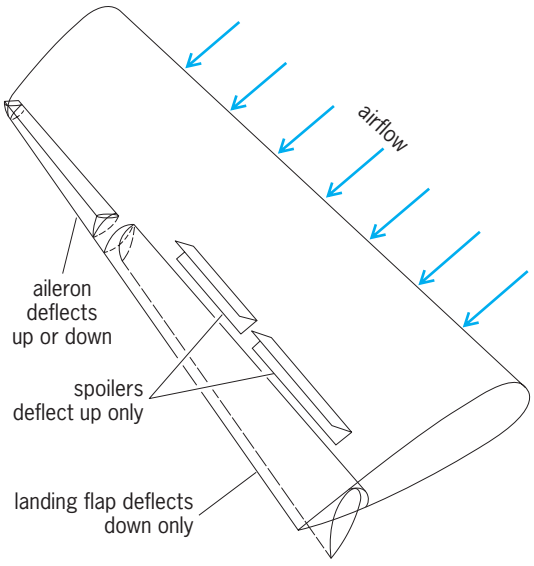


Fig. 18. Trailing-edge surfaces.

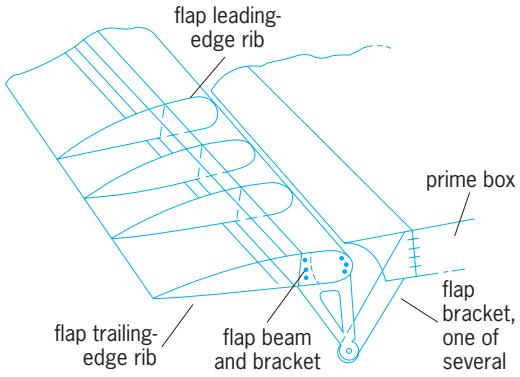


Fig. 19. Flap structure with skin removed. Although external hinges are shown, designs employing internal tracks and hinges are also used.

Fig. 20. Trailing edge with hinges for movable surfaces.

usual method is to build the fixed trailing-edge structure on the rear beam or rear spar of the wing with a master tool holding the locations of the hinge points for the movable surface (**Fig. 20**). The movable surfaces are then built from a matching hinge location tool. Thus, when the assembled surface is brought to the wing it is certain to fit without binding.

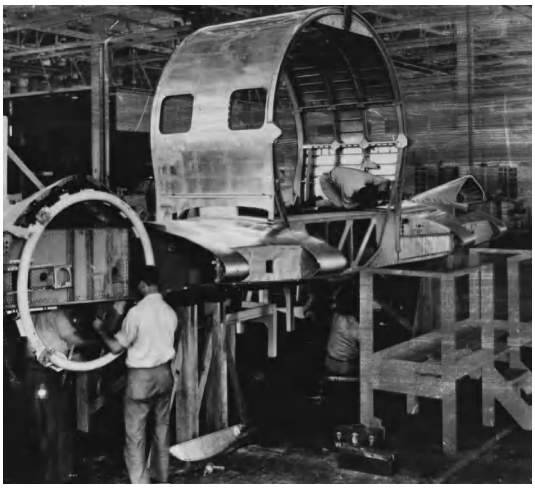


Fig. 21. Center wing-to-fuselage connection.

Wing-to-fuselage structure. The structural heart of the aircraft is the wing-to-fuselage joint (**Fig. 21**). This connection is usually the most complex in form and in analysis. Two major structural elements, the wing and the fuselage, with major loads running at right angles to one another, must be joined and analyzed for consistent deformations. The detail methods of indeterminate structural analyses are used. The calculations may require anywhere from 10 to 100 redundants. High-speed digital computers are a necessity.

The wing structures which have been described are attached to heavy aluminum ring frames in the fuselage; these rings distribute the wing loads to the fuselage skins. *See* FUSELAGE.

Assembly. The assembly of the wing can be traced by means of the illustrations. Figure 13 is representative of a main spanwise beam. To this, leading-edge structures, similar to that shown in Fig. 16, are added to produce a leading-edge and front-beam assembly. In the same manner, a beam and the fixed trailing-edge structures are joined together. Such an assembly is shown in Fig. 20. The rear beam and flap and aileron brackets are clearly indicated. These two major assemblies are then positioned in a major tool jig, and the intermediate beams, ribs, framing members, and top and bottom skins are progressively attached. Figure 15 shows such an assembly just before the attachment of the lower skins which close in the wing box. Figure 15 is an outer wing which eventually is attached as a unit to the center portion of the wing and fuselage. The center wing portion is shown in Fig. 21. The circular tubular structure at the left edge of Fig. 21 is part of the engine nacelle. *See* AIRFRAME.

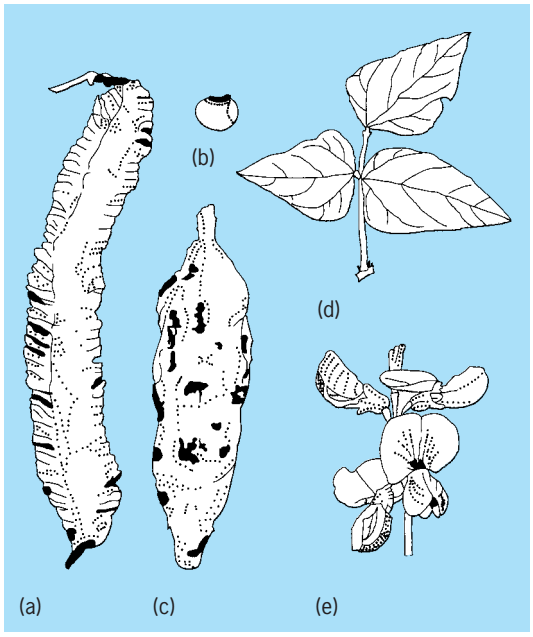
Harvey J. Hoge

Bibliography. American Society for Testing and Materials, *Damage Tolerance in Aircraft Structures*, STP 486, 1971; J. Cutler, *Understanding Aircraft Structures*, 3d ed., 1999; B. K. Donaldson, *Analysis of Aircraft Structures: An Introduction*, 1993; T. H. Megson, *Aircraft Structures for Engineering Students*, 2d ed., 1990; D. J. Peery and J. J. Azar, *Aircraft Structures*, 2d ed., 1982; S. P. Timoshenko and J. Gere, *Theory of Elastic Stability*, 2d ed., 1961.

Winged bean

A plant (*Psophocarpus tetragonolobus*), also known as four-cornered bean, asparagus pea, goa bean, and manila bean, in the family Leguminosae. It is a climbing perennial that is usually grown as an annual. It has been suggested that it originated either in East Africa or Southeast Asia, but there is more evidence to support an African origin. However, Southeast Asia and the highlands of Papua New Guinea represent two foci of its domestication.

Traditionally, winged bean is grown as a backyard vegetable in Southeast Asia and a few islands of the Pacific. It is grown as a field crop in Burma and Papua New Guinea. However, between 1980 and 1990 winged bean was introduced throughout the tropical world.



Edible parts of the winged bean: (*a*) pod, (*b*) seed, (*c*) root-tuber, (*d*) leaf, (*e*) flowers.

Nutritional value. Almost all parts of this plant are edible and are rich sources of protein. The green pods, tubers, and young leaves can be used as vegetables, and the flowers can be added to salads (see **illus.**). The dry seeds are similar to soybeans and can be used for extracting edible oil, feeding animals, and making milk and traditional Southeast Asian foods such as tempeh, tofu, and miso. Flour from the winged bean can also be used as a protein supplement in bread making.

The winged bean seed is nutritionally the most important product, containing 30–43% protein and 11–24% fat. The amino acid composition is similar to that of soybean, being low in sulfur-containing amino acids. The saturated/unsaturated fatty acid ratio is 1:3. The tocopherol content is high. The oil is refined easily and is reasonably stable. Phosphorus and zinc occur in significant quantities, and the vitamins thiamine and riboflavin are present in amounts comparable to that of other grain legumes. Tubers contain 8–10% protein on a fresh-weight basis, although the essential amino acids occur in low amounts. Immature green pods contain 1–3% protein and are rich sources of calcium, iron, and vitamin A. Leaves contain 5–7% protein on a fresh-weight basis and large amounts of vitamin A, vitamin C, and essential minerals.

Cultivation and harvesting. The winged bean is best adapted to the equatorial climate. Although it is grown at altitudes up to 6600 ft (2000 m) in Burma and Papua New Guinea, it does not tolerate frost. Short days (less than 12 h of sunlight) are necessary for both flower and tuber initiation. It can grow on a variety of soil types, but good drainage is necessary. It nodulates in association with the cowpea group of *Rhizobium* strains, which are widely distributed in the tropical soils. It responds to phosphorus fertilizers and also to potassium if phosphorus is not limiting.

In the tropics, sowing should be done at the commencement of the rainy season. Irrigation is necessary if a prolonged dry spell occurs before crop maturity. Weed control is needed in the first 4–6 weeks. Plants need to be supported by providing stakes or making a trellis; otherwise, the yield is seriously reduced. Fresh tender pods are harvested for vegetable when they are about 80% grown. Harvesting for seed is commenced as pods begin to dry out. This should be done periodically because if the pods are allowed to remain on the plant too long, pod splitting and seed shattering may occur. In the crop raised for tubers, pruning flowers and young pods increases the yield. Tubers are harvested by digging at the first sign of crop maturity.

Winged bean is a semidomesticated plant, and despite great potential it is a valuable source of food and feed mainly in subsistence farming. Plant breeding to evolve self-supporting growth habit, reduction in pod splitting and seed shattering, and synchronization of pod maturity will be necessary in order to develop winged bean varieties that can be grown broadly as an acreage crop.

Fresh harvested green pods keep poorly and should be marketed within 24 h. Tubers can be stored a little longer. The seed appears resistant to a number of storage insect pests, but seed viability declines quickly under tropical conditions. Seed intended for sowing should not be stored for longer than is necessary.

Diseases. A number of diseases and insect pests may limit winged bean yield. The most widespread and damaging disease appears to be false rust or orange gall (caused by *Synchytrium psophocarpi*). Dark leaf spot (caused by *Pseudocercospora psophocarpi*) and powdery mildew (caused by *Erysiphe cichoracearum*) are also important. Root knot nematodes (*Meloidoyne incognita*, *M. javanica*, and *M. arenaria*) cause galling of infected roots. The bean pod borer (*Maruca testulalis*) and other insect pests such as *Mylabris afzelli*, *M. pustulata*, *Heliothis armigera*, and *Icerya purchasi* have also been reported to cause damage to winged bean. *See* AGRICULTURAL SOIL AND CROP PRACTICES; BREEDING (PLANT); PLANT PATHOLOGY; ROSALES.

Tanveer N. Khan

Bibliography. T. N. Khan, *Winged Bean Production in the Tropics*, FAO Plant Prod. Protect. Pap. 38, Food and Agriculture Organization of the United Nations, 1982; R. J. Summerfield and E. H. Roberts (eds.), *Grain Legume Crops*, 1985.

Wire

A thread or slender rod of metal. Wire is usually circular in cross section and is flexible. If it is of such a diameter or composition that it is fairly stiff, it is termed rod. The wire may be of several small twisted or woven strands, but if used for lifting or in a structure, it is classed as cable. Wire may be used structurally in

tension, as in a suspension bridge, or as an electrical conductor, as in a power line. The working of metal into wire greatly increases its tensile strength. Thus, a cable of stranded small-diameter wires is stronger as well as more flexible than a corresponding solid rod. Wire may be treated or coated with various substances to protect it from corrosion or environmental influences. In addition, electrical conducting wire is usually covered with insulating material. *See* ELECTRICAL INSULATION; MAGNET WIRE.

Frank H. Rockett

Wire drawing

The reduction of the diameter of a metal rod or wire by pulling it through a die. The working region of dies are typically conical (**Fig. 1**). The tensile stress

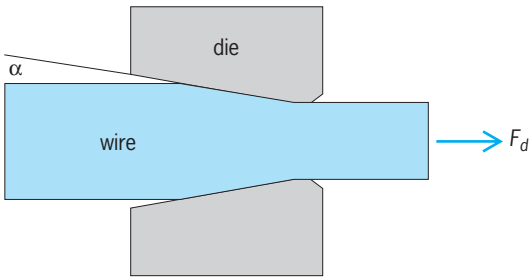


Fig. 1. Wire being drawn through a die.

on the drawn wire, that is, the drawing stress, must be less than the wire's yield strength. Otherwise the drawn section will yield and fail without pulling the undrawn wire though the die. Because of this limitation on the drawing stress, there is a maximum reduction that can be achieved in a single drawing pass. *See* STRESS AND STRAIN.

Efficiency. A simple analysis of drawing can be made with a work balance. The actual work per volume expended in drawing, w_a, can be divided into three terms: the ideal work per volume w_i, the frictional work per volume w_f, and the redundant work per volume w_r [Eq. (1)].

$$w_a = w_i + w_f + w_r \tag{1}$$

The total actual work W_a, is equal to F_dL, where F_d is the drawing force and L is the length drawn. Since the volume of the wire drawn is AL, where A is the cross-sectional area of the drawn wire, the work per volume w_a is equal to F_d/A, which is the stress on the drawn section σ_d [Eq. (2)].

$$w_a = \sigma_d \tag{2}$$

The term w_i is the work that would be required to make the same reduction in an imaginary tension test in which no necking occurs. This equals the area under the stress-strain curve. [Eq. (3)]. If work hardening is neglected, as is reasonable after a few passes, the integral $\int\sigma\, d\varepsilon$ may be approximated by $\sigma_{\text{av}}\Delta\varepsilon$, where σ_{av} is the average flow stress and $\Delta\varepsilon$ is the strain imposed during that pass.

$$w_i = \int \sigma\, d\varepsilon \tag{3}$$

The term w_f is simply the work against friction. It increases with decreasing die angle α, because the contact area between the wire and die increases with the imposed strain $\Delta\varepsilon$. For a constant coeficient of friction μ, Eq. (4) applies. If, instead of a constant friction coefficient, there is constant shear stress mk at the interface between the wire, the approximation is expressed by Eq. (5), where k is the shear strength of the wire.

$$w_f = w_i\mu\cot\alpha = \sigma_{\text{av}}\Delta\varepsilon\mu\cot\alpha \tag{4}$$

$$w_f = w_i m/(2\sin\alpha) = \sigma_{\text{av}}\Delta\varepsilon m/(2\sin\alpha) \tag{5}$$

The term w_r reflects the plastic work that is in excess of the ideal work required to produce the shape change. As a wire passes through a die, the surface layers are sheared relative to the center (**Fig. 2**). A simple upperbound model suggests that this term is independent of the strain $\Delta\varepsilon$ and depends only on flow stress and the die angle [Eq. (6)].

$$w_r = (2/3)\sigma_{\text{av}}\tan\alpha \tag{6}$$

It is often useful to define a mechanical efficiency η as in Eq. (7). Substituting Eqs. (3), (4), and (6), Eq. (8) can be written.

$$\eta = w_i/w_a \tag{7}$$

$$\begin{aligned}\eta &= \sigma_{\text{av}}\Delta\varepsilon/(\sigma_{\text{av}}\Delta\varepsilon + \sigma_{\text{av}}\Delta\varepsilon\mu\cot\alpha + (2/3)\sigma_{\text{av}}\tan\alpha)\\ &= 1/(1 + \mu\cot\alpha + (2/3)\Delta\varepsilon\tan\alpha)\end{aligned} \tag{8}$$

The variation of each of the work terms with die angle is shown schematically in **Fig. 3**. As the die angle α is increased, the redundant work increases and the frictional work decreases. The efficiency predicted by Eq. (8) is plotted in **Fig. 4** as a function of die angle for several reductions. Because Eq. (6) is an upperbound for w_r, a somewhat higher efficiency may be expected in practice. In any case, it is clear that there is an optimum die angle α for any given reduction and that α increases with increasing reduction.

The maximum reduction per pass is reached when the drawing stress σ_d equals the yield strength of the drawn wire. Except for the first few passes, work

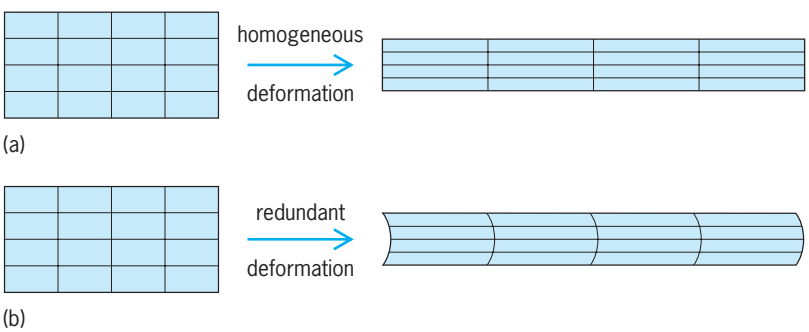


Fig. 2. Comparison of (*a*) ideal and (*b*) redundant deformation.

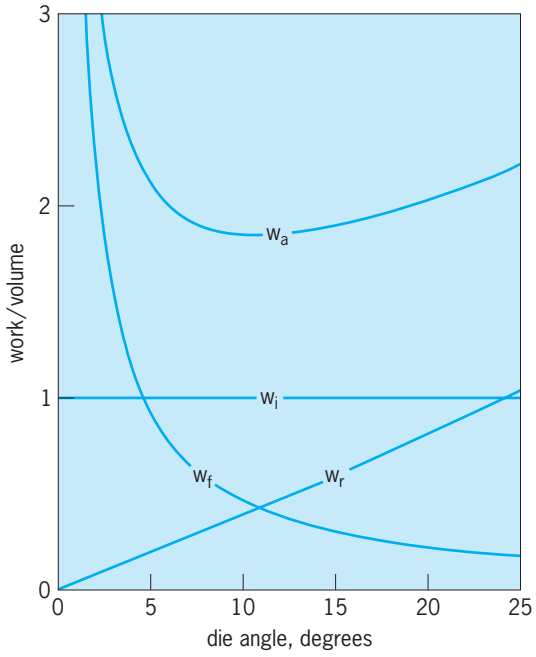


Fig. 3. Variation of the work terms with die angle according to Eqs. (1), (3), (4), and (6), with $\varepsilon = 0.3$ and $\eta = 0.08$.

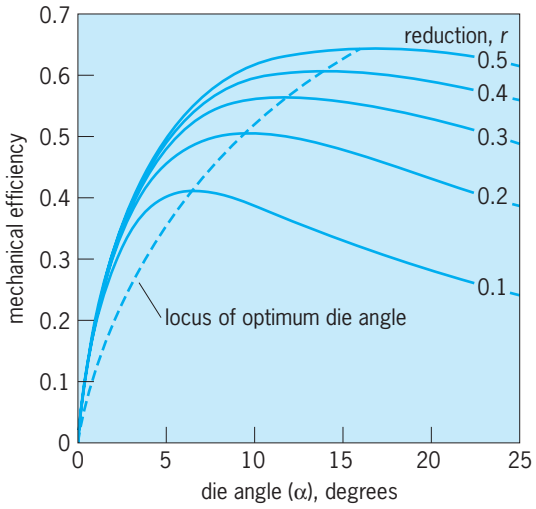


Fig. 4. Dependence of the efficiency on die angle for several reductions. A friction coefficient of $\mu = 0.08$ was assumed.

hardening can be neglected so the condition corresponds to $\sigma_d = \sigma_{av}$. Since $\sigma_d = w_a = (1/\eta)\sigma_{av}\Delta\varepsilon$, $\sigma_{av} = (1/\eta)\sigma_{av}\Delta\varepsilon$, or Eq. (9) applies. With a typical

$$\Delta\varepsilon = \eta \tag{9}$$

efficiency of 65%, $\Delta\varepsilon = 0.65$. This corresponds to an area reduction r_A as in Eq. (10), or 48%. The cor-

$$r_A = (A_o - A_f)/A_o = 1 - \exp(-\Delta\varepsilon) \tag{10}$$

responding diameter reduction $r_D = (D_o - D_f)/D_o = 1 - \exp(-\Delta\varepsilon/2)$, or 28%. In wire or rod production, multiple passes are used, without intermediate anneals, to produce fine wires. Reductions per pass are kept well below the maximum indicated by Eq. (9).

Uniformity. The homogeneity of deformation depends on the ratio Δ of the mean diameter of the wire in the deformation zone D_{av} to the contact length L between the wire and the die [Eq. (11)]. For

$$\Delta = D_{av}/L = 2\tan\alpha/(\Delta D/D_{av}) \tag{11}$$

$\Delta \leq 1$, the deformation is relatively uniform. Low die angles and high reductions per pass promote uniformity of product. However, with low reductions and high die angles, Δ may be greater than 1. Appreciable surface-to-center gradients of hardness and residual tensile stresses at the surface result when $\Delta > 1$. The inhomogeneity of deformation can be characterized by an inhomogeneity factor I.F. $= (H_s - H_c)/H_c$, where H_s and H_c are the surface and centerline hardnesses. For a material whose strain hardening is approximated by $\sigma = K\varepsilon^n$ an approximate relation is Eq. (12). **Figure 5** shows the predicted value of I.F.

$$\text{I.F.} = (1 + \tan\alpha)^n - 1 \tag{12}$$

according to Eq. (12) for drawn copper strip and wire together with experimental results. Equation (12) overpredicts I.F., but both the equation and the data indicate that the inhomogeneity increases with increasing die angle and decreasing reduction. *See* PLASTIC DEFORMATION OF METAL.

Crystallography. After large drawing reductions, wires or rods develop crystallographic textures or preferred orientations of grains. The textures are characteristic of the crystal structure of the metal. With face-centered cubic (fcc) metals, [100] and [111] directions become aligned with the wire axis. The relative amounts of the two orientations depends on the specific stacking fault energy of the metal. For aluminum and other metals with a very high stacking energy, the [111] component predominates, whereas [100] is the major component for silver. Alloys with even lower stacking fault energies have a predominant [111] component. For the latter materials, mechanical twinning is a dominant

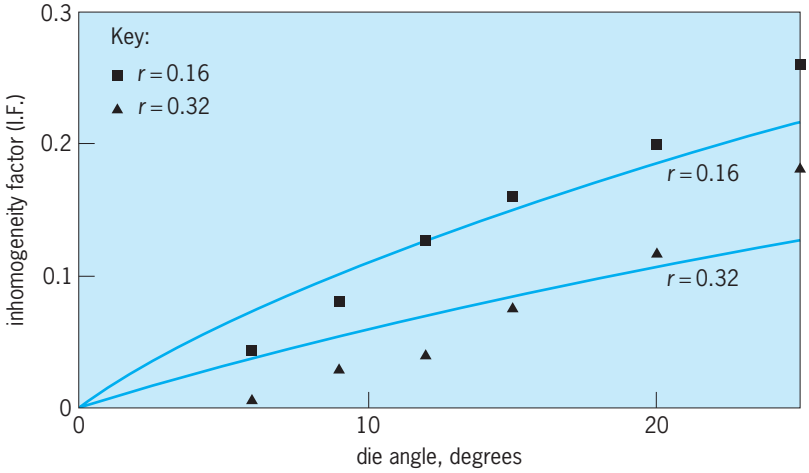


Fig. 5. Comparison of the predictions of I.F. from Eq. (12) with experiments on copper.

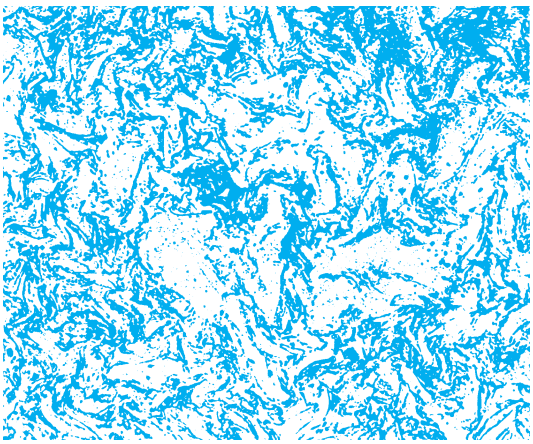


Fig. 6. Microstructure of a tungsten wire cold-drawn to an 87% reduction of area, viewed along the wire axis.

deformation mechanism. *See* ALLOY; CRYSTAL STRUCTURE; TWINNING (CRYSTALLOGRAPHY).

All body-centered cubic (bcc) metals develop a wire texture with only [110] aligned with the wire axis. Once this texture develops, the orientation of the slip systems is such that grains, as they elongate parallel to [110], thin in only the lateral [001] direction, with no strain in the lateral [110] direction. Neighboring grains must curl around each other to maintain compatibility [Eq. (4)]. This results in a characteristic microstructure (**Fig. 6**). Wire drawing of hexagonal close-packed metals forms a wire texture with the *c* axis normal to the wire axis. With this texture, grains thin in only the lateral direction normal to the *c* axis, so the microstructure formed is similar to that of bcc wires. William F. Hosford

Bibliography. W. F. Hosford, *The Mechanics of Crystals and Textured Polycrystals*, Oxford University Press, Oxford, 1993; E. M. Mielnik, *Metalworking Science and Engineering*, McGraw-Hill, New York, 1991.

Wireless fidelity (Wi-Fi)

In general, the wireless local area network (LAN) technology based on the Institute of Electrical and Electronics Engineers (IEEE) 802.11 standard. It enables computing devices to wirelessly exchange data with each other or with a wired network over a distance of up to about 300 ft (90 m), in a normal office environment, using unlicensed portions of the radio-frequency spectrum. Strictly speaking, Wi-Fi® technology refers to wireless LAN technologies that have passed interoperability tests designed by the Wi-Fi Alliance, an industry organization of vendors of 802.11 wireless LAN products. Such tests are focused on selected portions of the IEEE 802.11 standard and, sometimes, draft standards. *See* LOCAL-AREA NETWORKS.

The IEEE 802.11 standard consists of medium access control (MAC) specifications and physical layer (PHY) specifications. The MAC specifications define how a wireless LAN entity exchanges data with others using a shared wireless medium. The PHY specifications define the wireless signals that carry the exchanged data between wireless LAN entities and the wireless channels over which the wireless signals are transmitted. IEEE 802.11 was introduced in 1997 with a MAC specification and three PHY specifications. It has been amended with many extensions, including a MAC security specification 802.11i (2004) and three high-speed PHY specifications: 802.11b for 2.4 GHz (1999), 802.11a for 5 GHz (1999), and 802.11g for 2.4 GHz (2003). The IEEE 802.11 standard is still in fast evolution, as is Wi-Fi technology. *See* RADIO SPECTRUM ALLOCATIONS.

IEEE 802.11 network architecture. IEEE 802.11 defines two networking components, a station and an access point. A station has a MAC layer and a PHY layer; it is responsible for sending or receiving data to or from another station over the wireless medium, and does not relay data for any other station. An access point is a special station whose main task is to relay data for stations associated with it.

The data transmitted by the MAC layer are packed into frames. A frame consists of a header, a payload, and a tail. The header contains control information such as frame type, frame length, MAC addresses, and time window to be reserved for transmission of the frame. The payload can accommodate up to 2304 bytes of data, which may be encrypted. The tail is a 32-bit cyclical redundancy check (CRC) code that can tell a receiving MAC layer whether the received frame is corrupted.

IEEE 802.11 defines three wireless LAN architectures: independent basic service set (independent BSS), infrastructure BSS, and extended service set (ESS). An independent BSS (**Fig. 1***a*) is a wireless LAN consisting of stations only, and is usually set up on a temporary basis for a specific purpose. Every station competes for the wireless medium to send data frames to another station directly, and there is no guarantee that any two stations in the same independent BSS are in direct communication range. An infrastructure BSS (Fig. 1*b*) consists of an access point and a number of stations associated with it. The access point may be attached to a network of any kind, but most likely a wired LAN, called a distribution system (DS). Each station communicates only with the access point, and the access point relays data frames exchanged between these stations or between them and the DS. An ESS (Fig. 1*c*) consists of one or more infrastructure BSSs with all access points attached to a DS. If an ESS has two or more infrastructure BSSs, stations can utilize inter-BSS mobility by changing association from one access point to another without interrupting communications. *See* DATA COMMUNICATIONS.

IEEE 802.11 MAC. The IEEE 802.11 MAC provides three functions: fair access to the shared wireless medium with optional quality of service, reliable delivery of data frames over the wireless medium, and security protection.

Fair access. The basic medium access mechanism

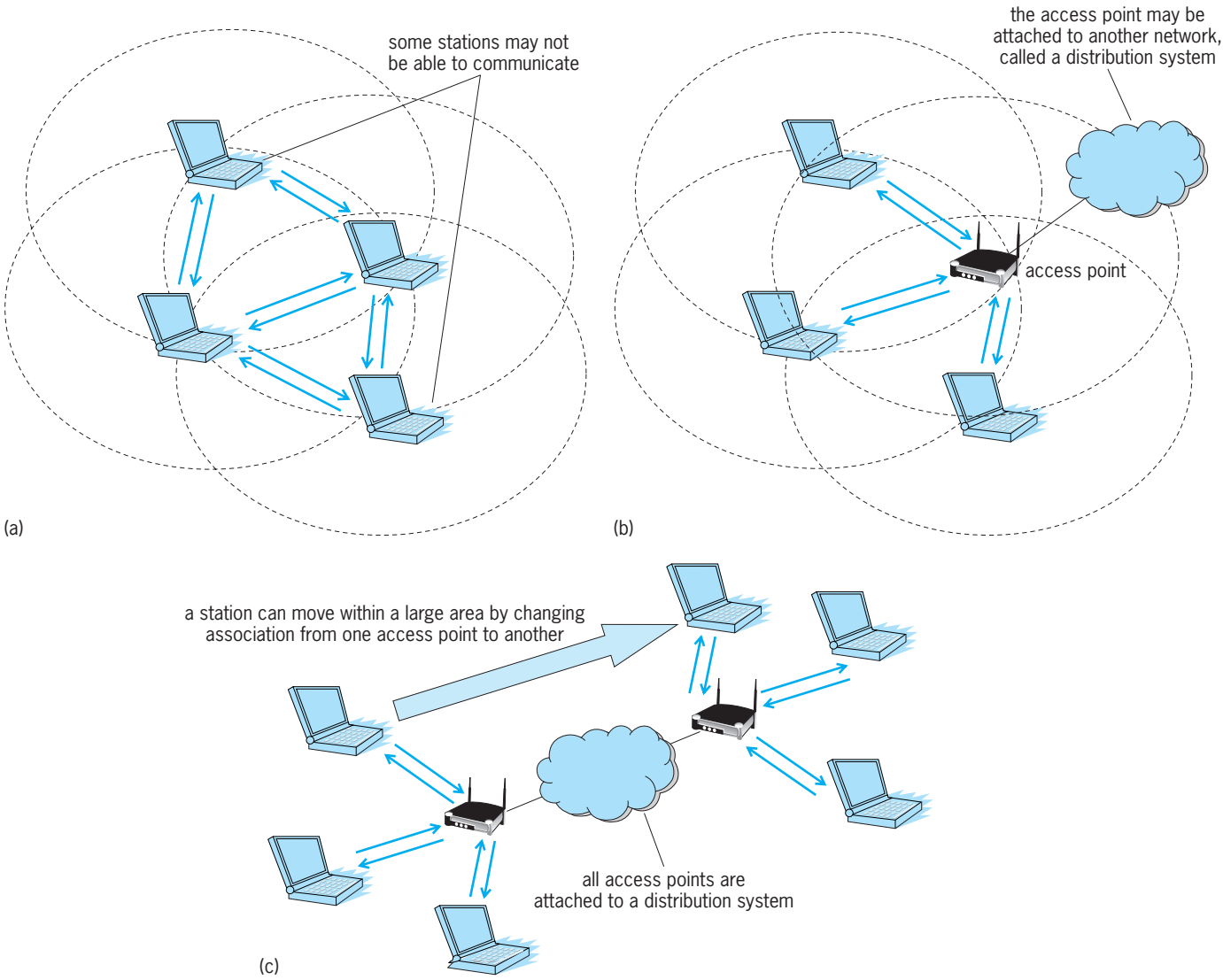


Fig. 1. Wireless LAN architectures. (*a*) Independent basic service set. (*b*) Infrastructure basic service set. (*c*) Extended service set.

employed by the IEEE 802.11 MAC is called carrier sense multiple access with collision avoidance (CSMA/CA). CSMA/CA is a wait-before-send protocol; that is, after the shared wireless medium becomes available, a station must wait a required period plus a back-off time before it can transmit a data frame, provided that the wireless medium is still available at the end of back off (**Fig. 2**). The back-off time is randomly selected by the station within a specified time period, called a contention window. CSMA/CA ensures that stations with the same contention window size and the same required waiting period have an equal probability to seize the shared wireless medium to transmit. It can reduce transmission collisions, thus achieving efficient utilization of the wireless medium, by randomizing the back-off time among station and enlarging the contention window size in case of transmission failure (which often signals congestion).

Based on CSMA/CA, the IEEE 802.11 MAC defines two medium access methods: distributed coordination function (DCF) and point coordination function (PCF). DCF assumes all stations use the same contention window range and the same required waiting period [either a DCF interframe space (DIFS) time or an extended interframe space (EIFS) time]. Since every station has an equal probability to transmit at every transmission attempt, DCF cannot support quality-of-service features such as traffic prioritization and guaranteed bandwidth.

PCF is an optional medium access method that can provide differentiated services for two classes of stations. It runs in conjunction with DCF and must be carried out by an access point. The access point maintains a list of stations that request higher quality of service. It periodically enters a contention-free period from DCF by transmitting a contention-free beacon frame, whose header announces the time window reserved for the forthcoming contention-free period. The access point transmits the beacon frame without back-off; it waits only a PCF interframe space (PIFS) time (shorter than DIFS and EIFS) after

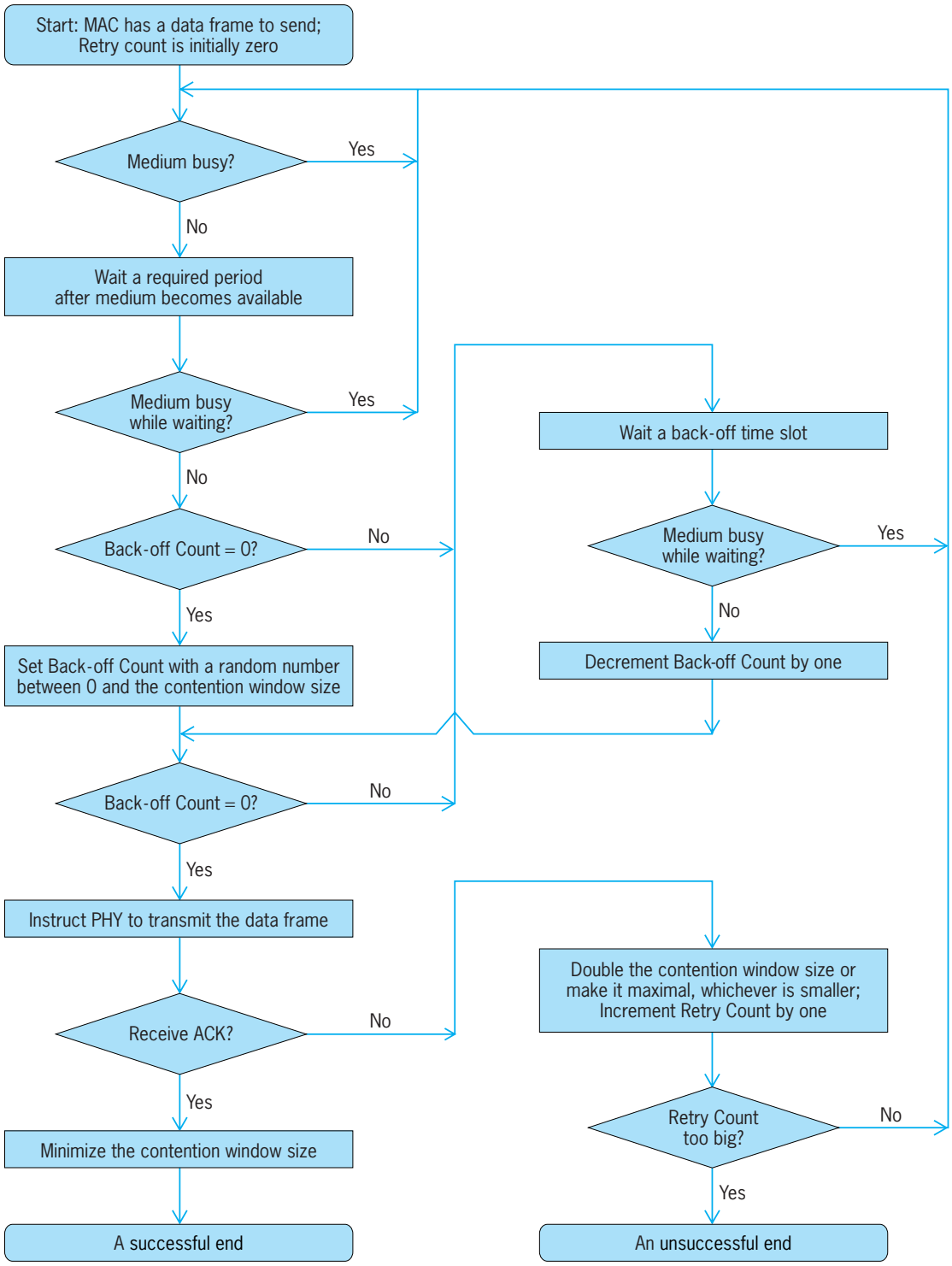


Fig. 2. Sending a data frame using carrier sense multiple access with collision avoidance (CSMA/CA).

the wireless medium becomes available. Therefore, none of the stations in the access point's radio range can compete for the wireless medium when the access point transmits the beacon frame, and they will not compete for the wireless medium during this period after receiving the beacon frame. The access point then polls stations on the list one by one and relays traffic for them with a guaranteed bandwidth and delay bound in this period.

To serve more complicated quality-of-service requirements, new medium access methods are developed in the 802.11e draft standard. In order to respond to market demand, the Wi-Fi Alliance started an interoperability test in 2004 for one of these methods, called EDCA in the 802.11e draft standard, or Wi-Fi multimedia (WMM) by the Wi-Fi Alliance. WMM provides differentiated services for four traffic classes. They are voice (highest priority),

video, best-effort data, and background data (lowest priority). WMM assigns a smaller contention window range and a shorter required waiting time to data frames with a higher priority. As a result, high-priority data frames have a greater probability to be transmitted than low-priority data frames at every transmission attempt. The bandwidth demanded by high-priority data frames can be guaranteed in a statistical sense regardless of how many low-priority data frames are waiting to be transmitted.

Reliable delivery. Since an 802.11 wireless LAN operates in unlicensed spectrum, the wireless medium could be very noisy due to the existence of other types of radio devices operating on the same band and not conforming to CSMA/CA. Thus, data frames can be frequently corrupted. In order to ensure reliable delivery of data frames over the noisy wireless medium, the IEEE 802.11 MAC employs a data frame exchange protocol for transmission of every data frame. It has a two-message form and a four-message form.

The two-message form is suitable for transmission of short data frames. It works as follows: (1) Station A sends a data frame using CSMA/CA to Station B, with the frame header announcing a reserved time window equal to the summation of the time needed to transmit the frame, a short interframe space (SIFS) time, and the time needed to transmit the expected acknowledge (ACK) frame. (2) If the data frame is correctly received, Station B sends a short ACK frame to Station A after waiting a SIFS time. (3) If Station A does not receive the ACK frame by the end of the reserved time window, it tries to resend the data frame using CSMA/CA. Note that the SIFS time is shorter than various required waiting time such as DIFS, EIFS, and PIFS time. This can prevent access points and stations that did not correctly receive the header of the data frame from interrupting the protocol.

The four-message form is designed to overcome the hidden node problem. A hidden node is a station that is outside the sending station's radio range but inside the receiving station's radio range (**Fig. 3**). Even if both the sending station and the hidden node run CSMA/CA, their data frames could still collide at the receiving station, because neither of them can detect the other's signal or time window reservation announcement in the frame header. The four-message data frame exchange protocol works as follows: (1) Station A sends a short request-to-send (RTS) frame using CSMA/CA to Station B, which tells the time window that must be reserved for the forthcoming protocol exchange. (2) Upon the reception of the RTS frame, Station B sends a short clear-to-send (CTS) frame to Station A after waiting only a SIFS time, which announces a time window reserved for the remained protocol exchanges to all stations in Station B's radio range, including hidden nodes to Station A, thus preventing them from sending during that period. (3) Upon the reception of the CTS frame, Station A sends the data frame to Station B after waiting a SIFS time. (4) If the data frame is correctly received, Station B sends a short ACK frame to Station A after waiting a SIFS time. (5) If Station A does not receive the CTS frame or the ACK frame on time, it tries to resend the RTS frame using CSMA/CA.

Security protection. Since anyone in the range of a wireless LAN can receive data frames transmitted over the wireless medium and can send data frames to stations or access points on the wireless LAN, authentication and data encryption must be implemented for access control and privacy. Initially, the IEEE 802.11 MAC defined the wired equivalent privacy (WEP) protocol to serve these needs. Unfortunately, WEP has significant security flaws. It allows authentication spoofing (falsification of a trusted identity). More seriously, an eavesdropper can break a 104-bit WEP key after collecting about 4 to 6 million encrypted data frames, which can be done in a time as short as a few hours. In order to replace WEP, the 802.11i standard has been developed. It specifies two authentication methods: preshared key (PSK) authentication (for simple wireless LAN configurations) and the IEEE 802.1x/extensible authentication protocol (EAP) authentication (for centralized access control on large wireless LANs); and two encryption methods: temporal key integrity protocol (TKIP) with message integrity check (MIC) and advanced encryption standard (AES).

TKIP uses the same stream cipher algorithm as WEP in data encryption, but corrects the security flaw in generating input to the stream cipher

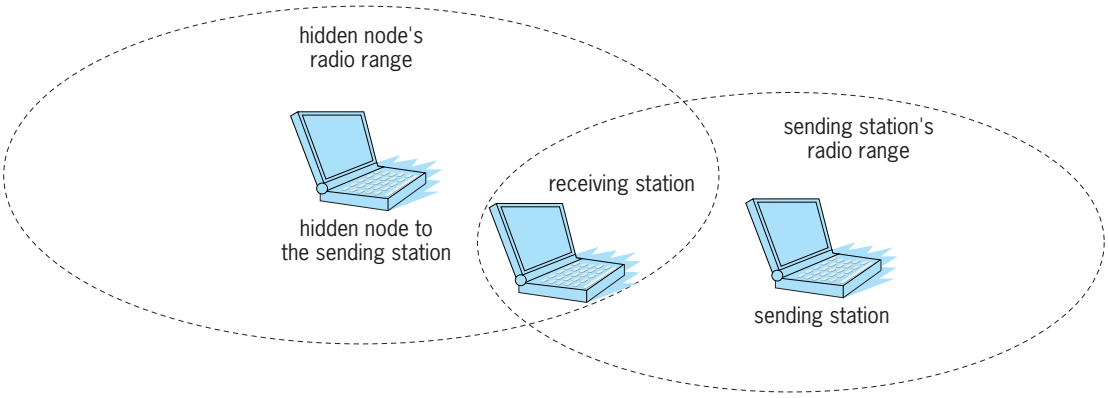


Fig. 3. Hidden node problem for a wireless network.

TABLE 1. The 2.4-GHZ frequency plan and power regulation applied to 802.11b

Channel numbers	Channel center frequencies, MHz	North America, 1000 mW (maximum output power)	ETSI (European Telecommunications Standards Institute) regulatory domain, 100 mW (effective isotropic radiation power)	Japan, 10 mW/MHz (220 mW maximum output power)
1	2412	X	X	X
2	2417	X	X	X
3	2422	X	X	X
4	2427	X	X	X
5	2432	X	X	X
6	2437	X	X	X
7	2442	X	X	X
8	2447	X	X	X
9	2452	X	X	X
10	2457	X	X	X
11	2462	X	X	X
12	2467		X	X
13	2472		X	X
14	2484			X

algorithm. It also introduces a MIC code associated with every encrypted data frame, such that the receiving station can verify authenticity of the frame. AES is a block cipher algorithm that demands significant computing power on stations, but it is considered much safer than the stream cipher algorithm used in TKIP. It combines encryption and MIC.

In the Wi-Fi Alliance's vocabulary, the wireless LAN security technology consisting of PSK authentication, the 802.1x/EAP authentication, and TKIP encryption with MIC is called Wi-Fi protected access (WPA); and that of PSK authentication, the 802.1x/EAP authentication, and the AES encryption is called WPA2. *See* COMPUTER SECURITY; CRYPTOGRAPHY.

IEEE 802.11 PHY. As of 2005 three 802.11 PHY had been put in use: 802.11b, 802.11a, and 802.11g, with 802.11b and 802.11g having a large installed base.

The 802.11b standard defines a direct sequence spread spectrum (DSSS) radio PHY for the 2.4-GHz industrial, scientific, and medical (ISM)

TABLE 2. The 5-GHz frequency plan and power regulation applied to 802.11a

		United States		CEPT (European Conference of Postal and Telecommunications Administrations) regulatory domain	
Channel numbers	Channel center frequencies, MHz	Channel allocation	Power regulation (maximum output power with up to 6 dBi antenna gain)	Channel allocation	Power regulation (effective isotropic radiation power)
36	5180	X	40 mW	X	200 mW
40	5200	X		X	
44	5220	X		X	
48	5240	X		X	
52	5260	X	200 mW	X	200 mW
56	5280	X		X	
60	5300	X		X	
64	5320	X		X	
100	5500			X	1000 mW
104	5520			X	
108	5540			X	
112	5560			X	
116	5580			X	
120	5600			X	
124	5620			X	
128	5640			X	
132	5660			X	
136	5680			X	
140	5700			X	
149	5745	X	800 mW		
153	5765	X			
157	5785	X			
161	5805	X			

band. The detailed worldwide frequency plan is shown in **Table 1**. In North America, there are 11 channels. Each channel spans 22 MHz; the center-frequency distance from that of adjacent channels is 5 MHz. Every channel partially overlaps on several adjacent channels, and thus there are effectively only three nonoverlapping channels available. The 802.11b standard specifies four data rates for the DSSS radio: 1, 2, 5.5, and 11 megabits per second (Mbps). The major interfering sources for 802.11b-based wireless LAN are microwave ovens, 2.4-GHz cordless phones, and Bluetooth® devices.

The 802.11a standard defines an orthogonal frequency division multiplexing (OFDM) radio PHY for the 5-GHz unlicensed national information infrastructure (U-NII) band. The detailed frequency plan for the United States and Europe is shown in **Table 2**. In the United States, there are 12 nonoverlapping channels available in three groups with different limits on transmitter power. Special spectrum issues such as power regulation and channel spacing for 802.11a wireless LAN operations in the 5-GHz band in Europe and Japan are specified in the 802.11h and 802.11j standards, respectively. The 802.11a standard defines eight data rates for the OFDM radio: 6, 9, 12, 18, 24, 36, 48, and 54 Mbps. Compared with 802.11b, 802.11a offers higher speed and more nonoverlapping channels but in general has a smaller operating range since 5-GHz wireless signals are attenuated faster than 2.4-GHz ones. For these reasons, 802.11a is more suitable than 802.11b for dense, high-speed installations.

The 802.11g standard defines an OFDM radio PHY that supports the same high data rates offered by 802.11a but remains compatible with the popular 802.11b. Specifically, the compatibility means that (1) every 802.11g station supports 802.11b PHY, so it can communicate with an 802.11b station using 802.11b; and (2) in an infrastructure BSS consisting of an 802.11g access point and a mixture of 802.11g and 802.11b stations, the 802.11g stations can communicate with the 802.11g access point using 802.11g for high speed, while 802.11b stations can communicate with the 802.11g access point using 802.11b, with no interference between 802.11g stations and 802.11b stations. Hui Luo

Bibliography. ANSI/IEEE Std 802.11, *Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications*, 1999; *Amendment to IEEE Std 802.11, 1999 Edition (Reaff 2003), IEEE Standard for Information technology—Telecommunications and information exchange between system—Local and metropolitan area networks—Specific requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications—Amendment 6: Medium Access Control (MAC) Security Enhancements*, 2004; IEEE Std 802.11a, *Supplement to IEEE Standard for Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: High-Speed Physical Layer in the 5 GHz Band*, 1999; IEEE Std 802.11b, *Supplement to IEEE Standard for Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Higher-Speed Physical Layer Extension in the 2.4 GHz Band*, 1999; IEEE Std 802.11g, *Supplement to IEEE Standard for Information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Specific requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Further Higher Data Rate Extension in the 2.4 GHz Band*, 2003; B. O'Hara and Al Petrick, *IEEE 802.11 Handbook: A Designer's Companion,* 2d ed., IEEE Press, 2005.

Wiring

A system of electric conductors, components, and apparatus for conveying electric power from a source to the point of use. In general, electric wiring for light and power must convey energy safely and reliably with low power losses, and must deliver it to the point of use in adequate quantity at rated voltage. To accomplish this, many types of electric wiring systems and components are used.

Electric wiring systems are designed to provide a practically constant voltage to the load within the capacity limits of the system. There are a few exceptions, notably series street-lighting circuits that operate at constant current.

In the United States, the methods and materials used in the wiring of buildings are governed as to minimum requirements by the National Electrical Code, municipal ordinances, and, in a few instances, state laws. The National Electrical Code is a standard approved by the American National Standards Institute (ANSI). Most materials used in wiring systems for light and power are tested and listed by Underwriters Laboratories, Inc. (UL). *See* ELECTRICAL CODES.

The building wiring system originates at a source of electric power, conventionally the distribution lines or network of an electric utility system. Power may also be supplied from a privately owned generating plant or, for emergency supply, a standby engine-generator or battery.

The connection from the supply to the building system through the metering devices, main disconnecting means, and main overcurrent protection constitute the "service entrance." The conductors, cables or busways, are known as service conductors. The switch and fuse or circuit breaker, serving as

the disconnecting means and the main overcurrent protection, are called the service equipment. Up to six individual switches or circuit breakers may be used for the service equipment to a single building, with the exclusion of certain specifically identified or required service disconnects (such as a fire pump disconnect).

As a rule, only one set of service conductors to a building is permitted, and conductors that run in parallel are considered one set. Large industrial plants, commercial buildings, and institutions are often served from more than one source. Separate service entrances are sometimes provided for emergency lighting, fire pumps, and similar loads.

Commercial and industrial systems. Three-phase wiring systems are generally used to conform to the supply systems. Energy is transformed to the desired voltage levels by a bank of three single-phase transformers or by a single three-phase transformer. The transformers may be connected in either a delta or Y configuration. In the delta configuration the ends of the transformer windings are connected together, and line conductors are connected to these points. A three-phase, three-wire system is thus formed, from which a single-phase line can be obtained from any two conductors. In the Y configuration one end of each transformer winding is connected to a common point, and line conductors are connected to the other ends of the transformer windings. This also forms a three-phase, three-wire system. A line wire is also often connected to the common point, forming a three-phase, four-wire system, from which single-phase circuits may be obtained between the common wire and any other. In typical three-phase transformers, the windings are internally connected at the factory in either delta-delta or delta-wye configurations.

Service provided at the primary voltage of the utility distribution system, typically 13,800 or 4160 V, is termed primary service. Service provided at secondary or utilization voltage, typically 120/208 or 277/480 V, is called secondary service.

Primary service. Service at primary voltage levels is often provided for large industrial, commercial, and institutional buildings, where the higher voltage can be used to advantage for power distribution within the buildings or utility rates are less expensive when service is taken at primary voltage.

Where primary service is provided, power is distributed at primary voltage from the main switchboard through feeders to load-center substations installed at appropriate locations throughout the building (**Fig. 1**). The load-center substation consists of a high-voltage disconnect switch, a set of transformers, and a low-voltage switchboard enclosed in a heavy sheet-metal housing.

In practice, several feeder arrangements are employed. These include (1) single primary: a single primary feeder serving several substations; (2) multiprimary: individual primary feeders to each substation; (3) loop primary: two primary feeders serving several substations, interconnected to form a ring or loop; and (4) primary selective: two primary feeders serving several substations, connectable to either feeder by a selector switch.

Wiring from the low-voltage-switchboard end of the substation to the load follows conventional utilization-voltage practice.

In large industrial plants, the secondary circuits of several substations are sometimes interconnected by feeder ties through switches or circuit breakers. The switches may be closed to form a network, if the transformers are suitable for parallel operation, or operated to transfer the load from one substation to another.

Enclosed bus-bar systems, called busways, are frequently used in the wiring system of industrial plants and large buildings. Busways are made and shipped in standard lengths with a wide variety of fittings. They are connected together on the job and installed as service-entrance conductors or feeders. One type of busway is designed to receive bus plugs at intervals along its length, and thus functions both as a feeder and as a distribution board or panelboard. The plugs, which tap the feeder, may contain fuses or circuit breakers, but the tap is often carried down to a more readily accessible overcurrent protective device. *See* BUS-BAR.

Secondary service. This type of service supplies power to the building at utilization voltage. Most

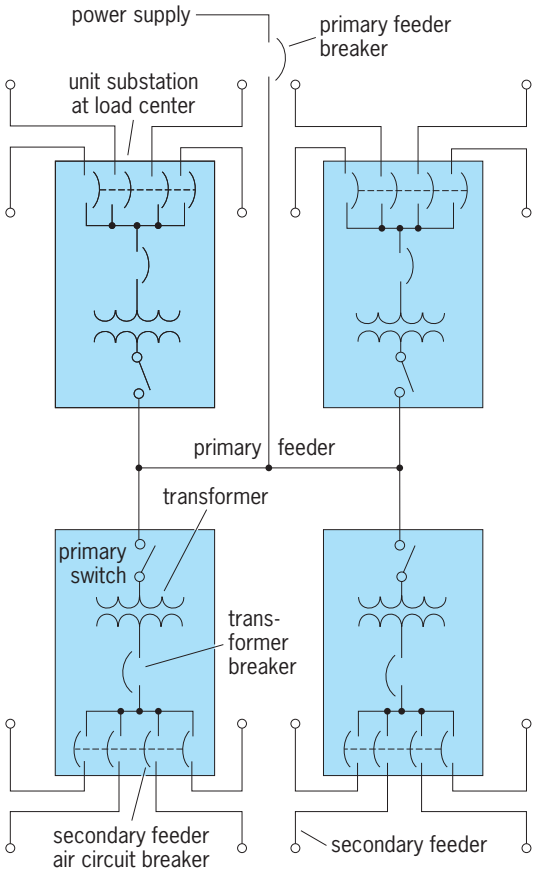


Fig. 1. Typical single primary distribution has a feeder from power supply to four load-center substations.

secondary services in the United States are 120/208 V, three-phase, four-wire, or 120/240 V, single-phase, three-wire serving both light and power. In some communities, separate light and power services are provided, typically 120/240 V, single-phase for lighting and 440 V, three-phase, delta for power.

Three-phase, four-wire services are almost universally Y configured, with the neutral tapped from the center of the Y. In some instances, a three-phase, four-wire service may be delta configured, with the neutral connected at the center tap of one phase. Such services provide typically 240 V, three-phase, three-wire for power and 120/240 V, single-phase, three-wire for lighting.

For relatively large buildings where the loads are predominantly fluorescent lighting and power (as for air conditioning), the service is often 277/480 V, three-phase, four-wire, supplying 480 V for power and 277 V, phase-to-neutral, for the lighting fixtures.

Distribution switchboards. From the service entrance, power is carried in feeders to the main switchboard, then to distribution panelboards (**Fig. 2**). Smaller feeders extend from the distribution panelboards to light and power panelboards. Branch circuits then carry power to the outlets serving the various lighting fixtures, plug receptacles, motors, or other utilization equipment. *See* BRANCH CIRCUIT.

The main distribution switchboard may also include the service equipment in its assembly. It consists of a group of switches and fuses or circuit breakers in a sheet-metal enclosure. It provides individual disconnecting and overcurrent protection for each feeder. Such equipment typically employs barriers between the service and distribution portions of the switchboard. In large buildings, additional distribution panelboards may be located at load centers.

Light and power panelboards provide individual disconnecting and overcurrent protection for the branch circuits. The circuit breakers of lighting panelboards are sometimes used as switches to operate the lighting circuits, when identified for such use.

Plug-receptacle power at 120/240 V, such as power for small appliances and business machines, in buildings provided with 277/480 V supply, is obtained from transformers. A feeder circuit from a 277/480 V distribution board or panelboard energizes the transformer primary. The secondary feeder serves a separate 120/240 V panelboard. The branch circuits serving the plug-receptacle outlets are conventional.

Wiring methods. Methods of wiring in common use for light and power circuits are as follows: (1) insulated wires and cables in metallic or nonmetallic raceways; (2) nonmetallic sheathed cables; (3) metallic armored cables; (4) busways; (5) copper-jacketed, mineral-insulated cables; (6) aluminum-sheathed cables; (7) nonmetallic sheathed and armored cables in cable support systems; and (8) open insulated wiring on solid insulators (knob and tube). *See* CONDUCTOR (ELECTRICITY).

Raceways in which insulated conductors may be installed are of several types: (1) rigid metal conduit; (2) electric metallic and nonmetallic tubing (EMT & ENT); (3) flexible metal and nonmetallic conduit; (4) liquid-tight flexible metal and nonmetallic conduit; (5) surface metal and nonmetallic raceway; (6) underfloor raceway; (7) cellular floor raceway; (8) rigid nonmetallic conduit; and (9) wireway.

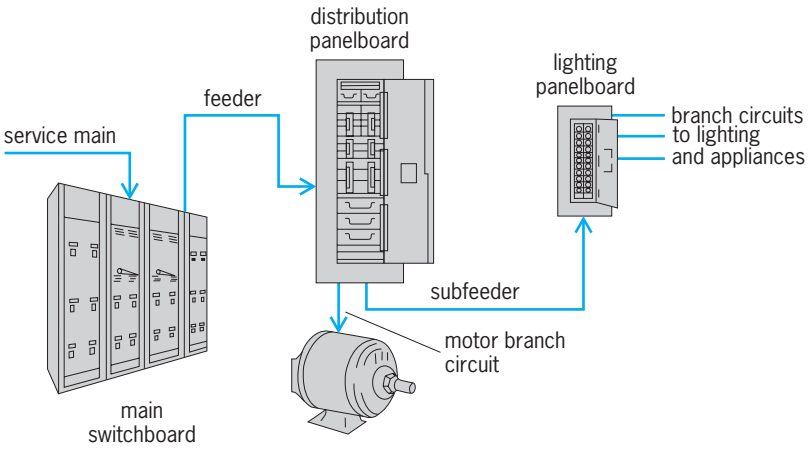


Fig. 2. Distribution wiring system has components to supply utilization voltage for several different types of load.

Insulated conductors may also be run in "cable tray," which is not considered a raceway but a cable support system.

The selection of the wiring method or methods is governed by a variety of considerations, which usually include code rules limiting the use of certain types of wiring materials; suitability for structural and environmental conditions; installation (exposed or concealed); accessibility for changes and alterations; and costs. Several methods may be employed together, for example, feeder busway risers in a multistory office building with the rest of the wiring in rigid conduit and underfloor raceways.

Circuit design. The design of a particular wiring system is developed by considering the various loads, establishing the branch-circuit and feeder requirements, and then determining the service-entrance requirements. Outlets for lighting fixtures, motors, portable appliances, and other utilization devices are indicated on the building plans, and the load requirement of each outlet is noted in watts or horsepower. Lighting fixtures and plug receptacles are then grouped on branch circuits and connections to the lighting panelboard indicated. Lighting and power panelboards are preferably located in the approximate center of the loads they serve; however, other considerations may require other locations. Panelboards in commercial and institutional buildings usually are located in corridors or electric closets. The size and number of panelboards are determined by the number of branch circuits to be served. *See* ELECTRIC POWER SYSTEMS.

Conductor sizes. The size of wires and cables used in electrical wiring systems is expressed in terms of the American Wire Gage (AWG), known also as the Brown and Sharpe (B&S) gage. Size designations run from No. 14, the smallest size commonly used in

wiring systems for light and power, to No. 4/0, the largest size in the gage. Sizes larger than No. 4/0 are designated by their cross-section areas expressed in circular mils. The largest size in practical usage is 2,000,000 circular mils (1013 mm^2). A circular mil is the area of a circle 0.001 in. (25.4 micrometers) in diameter.

Conductor capacity. The current-carrying capacity of wiring conductors is determined by the maximum insulation temperature that can be tolerated and the rate at which heat can be dissipated. All conductors offer some resistance to the flow of electric current. Consequently, heat is produced in the conductor by the flow of current I through its resistance R. The amount of heat is determined by the square of the current in amperes times the resistance in ohms (I^2R). Conductor heat is dispersed through the insulation and the surrounding raceway, and cable sheath or enclosure to the air. *See* ELECTRICAL INSULATION.

In practice, maximum current-carrying capacity of conductors is set forth in standard tables developed from laboratory tests and field experience. The National Electrical Code specifies the maximum current-carrying capacity of conductors. For any given size of conductor, the maximum capacity varies with the type of installation (in air or in raceways) and the maximum safe temperature of the insulation. Approved values are reduced for high ambient temperatures (greater than 86°F or 30°C) and for more than three current-carrying conductors in a single raceway or cable.

Feeders supplying several motors must be rated at not less than 125% of the full-load current of the largest motor plus 100% of the full-load currents of the remaining motors.

Feeders serving continuous loads that are likely to operate for 3 h or more (as office-building lighting) should not be loaded to more than 80% of rated capacity or have a capacity equal to 125% of the continuous load. The size of the feeder conductors will often be determined by the permissible voltage drop, which may require larger conductors than would be required by current-carrying capacity considerations alone.

Lighting loads. Lighting branch circuits may be loaded to 80% of circuit capacity. However, there is a reasonable probability that the lighting equipment will be added or replaced at some future time by equipment of higher output and greater load. Therefore, in modern practice, lighting branch circuits are loaded only to about 50% capacity, typically not more than 1200 W on a 20-A branch circuit.

Lighting branch circuits are usually rated at 20 A. Smaller 15-A branch circuits are used mostly in residences. Larger 30-, 40-, and 50-A branch circuits are limited to serving heavy-duty lampholders or appliances specially approved for connection to such circuits. The minimum conductor size for 20-A branch circuits is No. 12; however, on long runs larger conductors may be required to avoid excessive voltage drop. A common practice is to use No. 12 conductors between outlets and No. 10 conductors for the connection between the first outlet and the lighting panelboard.

Motor loads. These power loads are usually served by individual branch circuits, which must be rated at not less than 125% of the full-load rating of the motor.

Feeder and service-entrance design. The sum of the branch-circuit loads, including additional capacity for future load, determines the feeder load. The National Electrical Code provides a table of factors for various occupancies giving the minimum unit load per square foot and demand factors that may be applied. Demand factors are applied in installations where the loads are diversified and not likely to occur at one time. The number of feeders and their loads determine the number and size of distribution panelboard circuit elements required. The sum of the feeder loads determines the size and capacity of the service-entrance conductors and equipment.

Voltage drop. A drop in voltage along a conductor is a characteristic of all electric circuits. The voltage drop in a circuit causes the voltage at the load to be less than that applied to the circuit.

Wire or cable circuits are designed to carry a certain load but, whether the conductor be copper, aluminum, or other metal, the resistance characteristics impede the flow of current.

Since conductor resistance is proportional to conductor length, longer circuits are especially susceptible to excessive voltage drops. The amount of voltage drop E_D is determined by the current I in amperes times the resistance R in ohms ($E_D = IR$).

Good practice in circuit design dictates the following percentage values for maximum voltage drop: 5% from service entrance to any panelboard; 2% from panelboard to any outlet on branch circuit; 4% in feeders and 1% in branch circuit to motor; and 2% total in conductors to electric heating equipment.

The feeder voltage drop is calculated by formulas derived from Ohm's law and the resistance of the conductor. For three-wire, three-phase circuits (neglecting inductance), the voltage drop E_D is given in the equation $E_D = 1.732KIL/CM$, in which K is the resistance of a circular mil-foot of wire (for copper, 10.8 ohms), I is the current in amperes, L is the circuit length (source to load) in feet, and CM is the cross-section area of the conductor in circular mils.

Copper loss. This characteristic of a circuit is related to voltage drop. It is a power loss, designated as the I^2R loss, and is often expressed in percentage as the ratio of the wattage loss to the wattage delivered to the circuit.

Circuit protection. In wiring systems of high capacity (typically 1200 A and above) supplied from utility networks of large capability, overcurrent protective devices of high interrupting capacity are required. Circuit breakers of special design or current-limiting fuses are employed in such installations. In some cases, current-limiting reactors or bus-bar arrangements presenting appreciable reactance under short-circuit conditions are inserted in the service conductors. *See* CIRCUIT BREAKER; ELECTRIC

PROTECTIVE DEVICES; FUSE (ELECTRICITY).

Residential systems. For any house, the wiring system is an installed hookup of electrical conductors and components that provide for carrying electrical energy from the utility company lines to lighting fixtures, appliances, and receptacles. Every residential wiring system must conform in design and installation methods to the regulations of the National Electrical Code and any other codes that apply in a particular locality.

Every residential electrical system can be broken down into four basic categories: service entrance equipment; circuit wiring and raceways; outlet boxes and wiring devices; and fittings, connectors, and accessories.

Service entrance. As discussed above, the part of the wiring system that connects directly to the utility supply line is referred to as the service entrance. Depending upon the type of utility line serving the house, there are two basic types: overhead and underground. Service-entrance conductors are typically assemblies that consist of two or more insulated conductors combined with one bare conductor that is used as the grounded neutral, all encased in an outer nonmetallic jacket.

In the vast majority of cases, these conductors feed the service panelboard, which is an enclosure that is designed to simplify residential service-entrance installations carrying currents up to 200 A for 120/240 V, single-phase, three-wire systems. The service panelboard typically uses circuit breakers, although fuses, or a combination of fuses and circuit breakers, may be used in older installations to provide control and protection for the incoming service-entrance conductors and the branch-circuit wires, which are carried throughout the house to supply the lighting, appliances, and receptacles.

Circuit wiring and raceways. Branch circuits are run typically from the service panelboard in a residence to the various electrical outlets that must be supplied. Circuits to plug receptacle outlets and lighting fixture outlets generally are rated at 15 or 20 A and utilize two wires: one hot conductor and one neutral conductor. Circuits rated above 30 A are used only for supplying individual appliances, such as air conditioners, electric ranges, stoves, heaters, and clothes washers and dryers, and may include two hot conductors and a neutral.

A number of different types of cable are available for use in running branch circuits in the house and underground outdoors for lighting and receptacle circuits around yards, patios, and similar areas. The three most commonly used types are nonmetallic-sheathed cable (type NM), armored cable (type AC), and underground feeder cable (type UF).

Nonmetallic sheathed cable (type NM) is an assembly of two or more insulated conductors and usually a bare or an insulated ground conductor. For use in conformity with the National Electrical Code, any such cable assembly must be an approved type NM cable in sizes No. 14 to No. 2 American Wire Gage with copper conductors, and in sizes No. 12 to No. 2 with aluminum conductors. A corrosion-resistant type of nonmetallic sheathed cable (type NMC) is also available for use in damp or corrosive locations. Such conductors may be used in residential buildings of any construction type or height.

Armored cable (type AC) is a fabricated assembly of two or more insulated conductors in an interlocked, flexible metal jacket with a bare, internal, bonding conductor in intimate contact with the armor for its entire length.

Underground feeder cable is a factory-fabricated cable of one or more insulated conductors contained within an overall nonmetallic sheath that is suitable for direct burial. Typical applications include primarily lighting, as well as power loads other than swimming pools, hot tubs, fountains, and the like.

Outlet boxes and wiring devices. A wide variety of small boxes are employed at various points in the wiring system where electrical power is supplied to lighting fixtures, motors, fixed appliances, and cord-connected devices and appliances. Such boxes are available in both metallic and nonmetallic types. Metallic boxes are used in all wiring systems, while the nonmetallic types are generally used with type NM cable systems.

Wiring boxes are commonly divided into several categories according to construction and application, as follows.

Octagonal and round outlet boxes are typically used for ceiling outlets.

Oblong and square boxes are commonly referred to as gang boxes because they permit the mounting of a number of standard wiring devices in a single box. A four-gang box, for example, will accommodate four standard wiring devices mounted side by side. Such boxes usually employ one of a variety of covers.

Sectional switch boxes have removable sides to allow adding boxes together, or ganging of boxes, to accommodate installation of several wiring devices side by side at one location. These boxes do not require covers; they accommodate standard flush wiring devices with standard flush wall plates.

Utility boxes are all-purpose boxes designed for surface mounting.

Wiring devices encompass the wide range of types and sizes of convenience receptacles, wall switches, cord plugs, and connectors.

Convenience receptacles are made in many types and sizes to provide for the connection of supply cords from lighting units, motor-operated devices, and other appliances. Such devices are classified according to the number and arrangement of the contact slots and the type of mounting, which are based on their voltage and current rating. *See* ELECTRICAL CONNECTOR.

Standard receptacles are made in single and duplex (two receptacle outlets on a single yoke) types for use in a single gang box. Interchangeable-type receptacles permit different combinations of receptacles, switches, and pilot lights to be assembled as desired. Up to four devices may be installed in a single gang

box on a single strap. Combination devices are similar to the interchangeable type in that a combination of switches, receptacles, and pilot lights, if desired, can be installed as a single-device gang box. However, combination devices are not field-assembled, but are manufactured in the various configurations and, as with other standard switches and receptacle, are enclosed by molded bodies. Additionally, many so-called pilot devices have been replaced with integral light-emitting diode (LED) indicators.

Receptacle outlets installed on 15- and 20-A branch circuits must be of the grounding type to ensure compliance with the National Electrical Code. Replacements for non-grounding-type receptacles in existing residences must be nongrounding receptacles or ground-fault circuit-interrupter-type (GFCI) receptacles, or nongrounding types provided with upstream GFCI protection. Additionally, ground-fault protection must be provided for receptacles installed in the bathroom, the kitchen countertop, outdoor locations, and where they are readily accessible or not dedicated to specific appliances in attached or detached garages and basements. The required ground-fault protection may be achieved with a Class A ground-fault circuit-interrupter breakers or receptacles. *See* GROUNDING.

Wiring devices also include a range of small switches. Such switches are used in nearly every type of interior wiring system to control branch-circuit equipment (lighting units or appliances). Generally, such switches operate on a snap-action principle, with a toggle handle, push button, or rotary handle operating a contact mechanism that gives definite on and off positions.

Fittings, connectors, and accessories. In electrical work, a wide variety of small hardware devices are used to provide complete installations of conductors in conduit or cable systems. These devices include mechanical connecting and fastening items that complete the raceway system, and items that provide for the connection of current-carrying conductors.

Joseph F. McPartland; Brian J. McPartland

Bibliography. T. Croft and W. I. Summers (eds.), *American Electricians Handbook*, 14th ed., 2002; D. G. Fink and H. W. Beaty (eds.), *Standard Handbook for Electrical Engineers*, 14th ed., 2000; Institute of Electrical and Electronics Engineers, *The Recommended Practice for Electrical Power Systems in Commercial Buildings*, Std. 241, 1990; J. F. McPartland, *National Electrical Code Handbook*, 25th ed., 2005; J. F. McPartland and B. J. McPartland, *Handbook of Practical Electrical Design*, 3d ed., 1999; National Fire Protection Association, *National Electrical Code 2005*, 2005; H. P. Richter and W. C. Schwan, *Practical Electrical Wiring*, 19th ed., 2005.

Wiring diagram

A drawing illustrating electrical and mechanical relationships between parts on a component that require interconnection by electrical wiring.

A wiring diagram is distingushed from an electrical schematic in that the arrangement of the schematic bears no necessary relationship to the mechanical arrangement of the electrical elements in the component. The wiring diagram provides an accurate picture of how the wiring on the elements and between them should appear in order that the electrical wiring technician can install the wiring in the manner that will best contribute to the performance of the device.

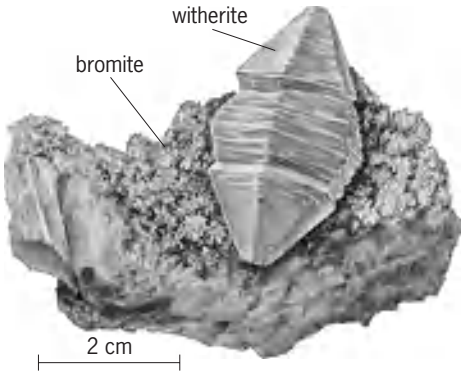
The degree of symbolism used in a wiring diagram depends on the extent of standardization in the particular field. For example, in telephone switchboard wiring, which consists of many standardized repetitive operations, extensive symbolism is used. When the exact physical location of wiring is important, as in radio-frequency devices where electromagnetic and electrostatic coupling between wires is appreciable, the diagram can be quite pictorial.

Wiring diagrams also include such information as type of wire, color coding, methods of wire termination, and methods of wire and cable clamping. *See* SCHEMATIC DRAWING.

Robert W. Mann

Witherite

The mineral form of barium carbonate. Witherite has orthorhombic symmetry and the aragonite structure type. Crystals, often twinned, may appear hexagonal in outline (see **illus.**). It may be white or gray with yellow, brown, or green tints. Its hardness is 3.5 and its specific gravity 4.3.



Large crystal with bromite from Fallowfield, Northumberland, England. (*Specimen from Department of Geology, Bryn Mawr College*)

Witherite may be found in veins with barite and galena. It is found in many places in Europe, and large crystals occur at Rosiclare, Illinois. *See* BARIUM; CARBONATE MINERALS.

Robert I. Harker

Wolf-Rayet star

A type of hot, luminous star that is distinguished by its extremely dense and fast wind. The spectacularly bright, discrete bands of atomic emission from these winds greatly facilitated their discovery with the

aid of a visual spectroscope by the French astronomers Charles Wolf and Georges Rayet at the Paris Observatory in 1867. *See* ASTRONOMICAL SPECTROSCOPY.

The Wolf-Rayet phenomenon includes a typical phase in the advanced evolution of a massive star (about 20–100 times the Sun's mass at birth), or sometimes (15% of cases) among lower-mass stars in the planetary nebula stage. Current evidence suggests that most massive Wolf-Rayet stars are the compact, hot cores remaining after most of the initial hydrogen-rich material in the central region of a massive star has been fused via nuclear burning to heavier elements. The surrounding unburnt, hydrogen-rich layers have been mostly removed by the efficient stellar wind. In a few rare cases, the Wolf-Rayet phenomenon is seen in the most massive, thus luminous, stars on the main sequence, even before central hydrogen-burning is completed (**Fig. 1**). *See* PLANETARY NEBULA.

The first fusion process in the core of a massive star involves the conversion of hydrogen into helium via the carbon-nitrogen-oxygen (CNO) cycle, in which helium and nitrogen are enhanced at the expense of the initially abundant hydrogen and traces of carbon and oxygen. When such fusion products are visible in the winds, a WN-type Wolf-Rayet star is seen (**Fig. 2**), whose spectrum is dominated by Doppler-broadened atomic lines of helium and nitrogen in various stages of ionization. Later, when the second fusion process occurs, helium is converted mainly into carbon and oxygen, with nitrogen being virtually destroyed. A WC-type Wolf-Rayet star is then seen with lines mainly of carbon and helium. A brief oxygen-rich phase may occur (WO) toward the end of the helium-burning phase, after which all subsequent exotic nuclear-burning phases are so rapid that they remain hidden by the much more slowly changing stellar surface. At that point, it is believed that the Wolf-Rayet star will explode as a supernova, resulting in the collapse to a black hole in most cases and an associated gamma-ray burst in very rare instances. *See* BLACK HOLE; DOPPLER EFFECT; GAMMA-RAY BURSTS; NUCLEOSYNTHESIS; SUPERNOVA.

Beneath the dense winds that often hide the stellar surface, massive Wolf-Rayet stars have surface temperatures ranging from 30,000 to 150,000 K (54,000 to 270,000°F), radii of 1 to 15 solar units, and luminosities of 10^5 to 10^6 times that of the Sun. They are losing matter at a rate that is typically 10^9 times that of the Sun's wind, at speeds ranging from 1000 to 3000 km/s (600 to 1800 mi/s), clearly exceeding the minimum speed required to escape from the star. *See* SOLAR WIND.

Although massive Wolf-Rayet stars appear to be rare (only about 250 are known so far in the Milky Way Galaxy, out of an estimated total population of 1000–2000), all massive stars likely pass through a Wolf-Rayet stage toward the end of their relatively short lives. The Wolf-Rayet phase lasts at most about 10% of the total lifetime of a massive star, the latter being some 2–6 million years. *See* STAR; STELLAR EVOLUTION.

Anthony F. J. Moffat

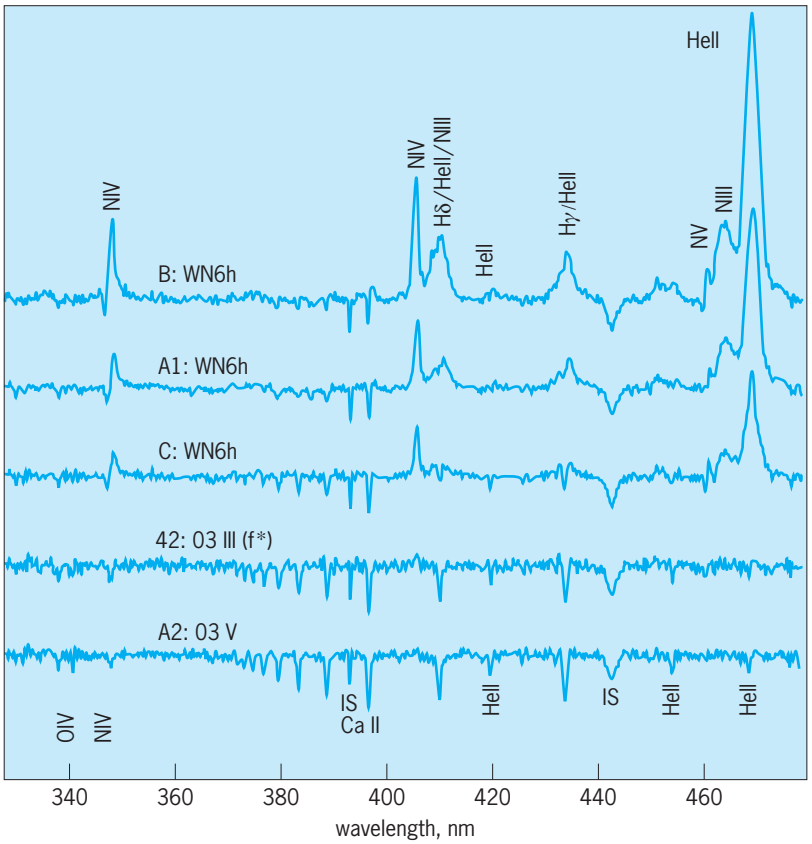


Fig. 1. Brightness-sorted montage of Hubble Space Telescope spectra for the 5 most luminous stars in the core of the extremely dense, young galactic cluster NGC 3603, showing its three hydrogen-rich, hydrogen-burning Wolf-Rayet stars, A1, B, and C. These three stars display strong Wolf-Rayet–like emission lines compared to the other normal, fainter O-type stars. (*From L. Drissen, The stellar content of NGC 3603, in K. A. van der Hucht, G. Koenigsberger, and P. R. J. Eenens, eds., Wolf-Rayet Phenomena in Massive Stars and Starburst Galaxies, International Astronomical Union Symposium 193, p. 403, Astronomical Society of the Pacific, 1999*)

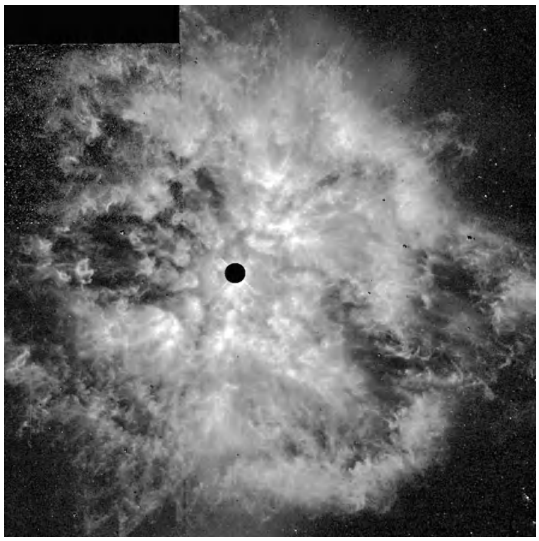


Fig. 2. Hubble Space Telescope image in the light of hydrogen-alpha of the Wolf-Rayet star WR 124 (blotted out in the center) and its surroundings. This is a particularly interesting fast-moving, hyrdogen-poor WN star, whose fast wind is catching up and strongly interacting with the interstellar medium and the slow wind of its immediate progenitor, a so-called luminous blue variable star or a red supergiant. (*From Y. Grosdidier et al., HST WFPC2/Hα imagery of the nebula M1-67: A clumpy LBV wind imprinting itself on the nebular structure?, Astrophys. J., 506:L127–L131, 1998*)

Bibliography. H. J. G. L. M. Lamers and J. P. Cassinelli, *Introduction to Stellar Winds*, Cambridge University Press, 1999; A. Maeder and P. S. Conti, Massive star populations in nearby galaxies, in G. Burbidge (ed.), *Annu. Rev. Astron. Astrophys.*, 32:227–275, 1994; K. A. van der Hucht, The VIIth Catalogue of galactic Wolf-Rayet stars, *New Astron. Rev.*, 45(3):135-232, 2001; K. A. van der Hucht, A. Herrero, and C. Esteban (eds.), *A Massive Star Odyssey: From Main Sequence to Supernova, International Astronomical Union Symposium 212*, Astronomical Society of the Pacific, 2003; K. A. van der Hucht, G. Koenigsberger, and P. R. J. Eenens (eds.), *Wolf-Rayet Phenomena in Massive Stars and Starburst Galaxies, International Astronomical Union Symposium 193*, Astronomical Society of the Pacific, 1999; J. M. Vreux et al. (eds.), *Wolf-Rayet Stars in the Framework of Stellar Evolution, 33d Liege International Astrophysical Colloquium*, 1996.

Wolframite

A mineral with chemical composition $(Fe,Mn)WO_4$, intermediate between ferberite, the iron tungstate, and huebnerite, the manganese tungstate, which form a complete solid solution series. *See* HUEBNERITE.

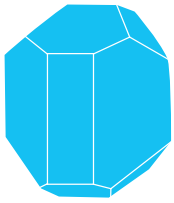
Wolframite occurs commonly in short, brownish-black, monoclinic, prismatic, bladed crystals (see **illus.**). It is difficult to dissolve in acids, but wolframite high in iron fuses readily to a magnetic globule.

Wolframite is probably the most important tungsten mineral. A quick and easy test for tungsten is to fuse the mineral powder with charcoal and sodium carbonate and boil the residue in hydrochloric acid with a few grains of granulated metallic tin. The presence of tungsten gives the solution a prussian-blue color. The major use of tungsten is in making ferrous (steel) alloys, nonferrous alloys, tungsten carbide, metallic tungsten, and tungsten chemicals.

Wolframite is found associated with quartz in veins in the peripheral areas of granitic bodies. It is also found in veins associated with sulfide minerals such as pyrite, chalcopyrite, arsenopyrite, and bismuthinite, together with cassiterite, molybdenite, hematite, magnetite, tourmaline, and apatite. It also occurs in placers.

China is the major producer of wolframite. Tungsten minerals of the wolframite series occur in many areas of the western United States; the major producing district is Boulder and northern Gilpin counties in Colorado. *See* TUNGSTEN.

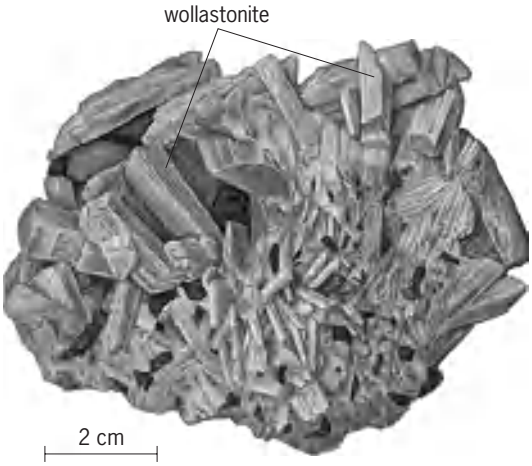
Edward C. T. Chao



Wolframite crystal habit. (*After C. Klein and C. S. Hurlbut, Jr., Manual of Mineralogy, 21st ed., John Wiley and Sons, 1993*)

Wollastonite

A mineral inosilicate with composition $CaSiO_3$. It crystallizes in the triclinic system in tabular crystals (see **illus.**). More commonly it is massive, or in cleavable to fibrous aggregates. There are two good cleavages parallel to the front and basal pinacoids yielding elongated cleavage fragments. Hardness is 5–5½ on Mohs scale; specific gravity is 2.85. On the cleavages the luster is pearly or silky; the color is white to gray. Wollastonite is the most common of three polymorphic forms of $CaSiO_3$, the other two being pseudowollastonite and parawollastonite. Pseudowollastonite, a high-temperature triclinic form, is very uncommon in rocks but may be a constituent of synthetic $CaO\text{-}SiO_2$ systems and of slags and glasses. Parawollastonite, a monoclinic form, is only rarely found in Ca-rich rocks. Wollastonite, by far the most common polymorph, occurs abundantly in impure limestones that have undergone contact metamorphism. Resulting assemblages may consist of calcite-diopside-wollastonite with variable amounts of tremolite, clinozoisite, and grossularite. Wollastonite occurs sporadically in regionally metamorphosed calcareous sediments as well. It is found in large masses in the Black Forest of Germany; Brittany, France; Chiapas, Mexico; and Willsboro, New York, where it is mined as a ceramic material. *See* SILICATE MINERALS.



Triclinic tabular crystals embedded in limestone, Santa Fe, Chiapas, Mexico. (*Specimen from Department of Geology, Bryn Mawr College*)

Cornelius S. Hurlbut, Jr.

Wolverine

The largest member of the family Mustelidae, to which weasels, mink, fishers, martens, ferrets, badgers, and otters belong.

Morphology. The powerfully built wolverine (*Gulo gulo*) has an overall shaggy appearance (see **illustration**). The heavy body is covered with long guard hairs that overlie a coarse, dense, kinky, woolly underfur. The general coloration is blackish brown. A light brown band extends along each side of the body from shoulder to rump before merging at the base of the tail. Most wolverines have a light silver facial mask. The large head is broad and somewhat rounded with small, wide-set eyes and short, rounded, well-furred ears. The legs, feet, and most of the relatively short, shaggy tail are dark. The legs are massive, stocky, and powerful, and the feet are large. The head and tail are carried lower than the somewhat arched back, and the lumbering plantigrade gait gives the impression of clumsiness. The palms and soles are densely haired. The claws are pale and sharply curved. A pair of anal scent glands are present which secrete a yellowish-brown fluid. The dental formula is I 3/3, C 1/1, PM 4/4, M $1/2 \times 2 = 38$ teeth. Adults have a head and body length of 650–1050 mm (25–41 in.), a tail length of 170–260 mm (6.5–10 in.), and weigh 7–32 kg (15–70 lb). Males are generally 25–30% heavier and 10% greater in most body measurements than females of similar age. *See* DENTITION.

Distribution and ecology. The wolverine has a circumpolar distribution, occurring in the tundra and taiga zones throughout northern Europe, Asia, and North America. It is uncommon in the United States and most of Canada and is very rare in Scandinavia. It is active all year and may be found in forests, mountains, or open plains. In mountainous areas, it moves to lower elevations during winter. Nests of grass and leaves are constructed inside the den, which may be in a cave or rock crevice, beneath a fallen tree, or in a burrow made by another animal. Wolverines are primarily terrestrial, but they can also climb trees and are excellent swimmers. These mammals have a keen sense of smell, but they lack keen senses of vision and hearing. Wolverines are solitary wanderers except during the breeding season. They occur at relatively low population densities and are territorial. They have enormous home ranges; average yearly home ranges in Montana are 422 km^2 for males and 388 km^2 for females. In south-central Alaska, the average home range is 535 km^2 for males and 105 km^2 for females with young.

Mating and reproduction. Although mating occurs from April to July, wolverines exhibit delayed implantation—that is, implantation of the fertilized eggs does not occur until November to March. Although total gestation may last from 215 to 272 days, active gestation (after implantation) is only 30 to 40 days. Births occur from January to April. Females produce litters of two to four about every 2 years. Young wolverines nurse for 8–10 weeks, leave their mother in the autumn, and reach sexual maturity in their second or third year of life. Longevity in the wild is approximately 8 to 10 years; however, it is possible for captive wolverines to live more than 17 years.

Dietary habits. Food consists of carrion (reindeer, red deer, moose, caribou), eggs of ground-nesting birds, lemmings, and berries. The tenacity and strength of the wolverine reportedly allow it to drive bears and cougars from their kills. Wolverines will attack large mammals when they are sick or trapped in deep snow; their large feet act like snowshoes, spreading the animal's weight so that it can run over soft, deep snow that slows down its heavy, hoofed prey.

Wolverine (*Gulo gulo*) in Kalispell, Montana. (*Photo by Gerald and Bliff Corsi, © 1999 California Academy of Sciences*)

Status. Humans, through direct persecution, deforestation, encroachment, and urbanization, are the only important enemy of the wolverine. It has been classified as vulnerable by the International Union for the Conservation of Nature (IUCN) and has been designated as endangered in eastern Canada. *See* BADGER; CARNIVORA; FERRET; FISHER; MARTEN; MINK; OTTER; WEASEL.

Donald W. Linzey

Bibliography. D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford Limited, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Wood anatomy

Wood is composed mostly of hollow, elongated, spindle-shaped cells that are arranged parallel to each other along the trunk of a tree. The characteristics of these fibrous cells and their arrangement affect strength properties, appearance, resistance to penetration by water and chemicals, resistance to decay, and many other properties.

Gross features of wood. Just under the bark of a tree is a thin layer of cells, not visible to the naked eye, called the cambium. Here cells divide and eventually differentiate to form bark tissue to the outside of the cambium and wood or xylem tissue to the inside. This newly formed wood (termed sapwood) contains many living cells and conducts sap upward in the tree. Eventually, the inner sapwood cells become inactive and are transformed into heartwood. This transformation is often accompanied by the formation of extractives that darken the wood, make it less porous, and sometimes provide more resistance to decay. The center of the trunk is the pith, the soft

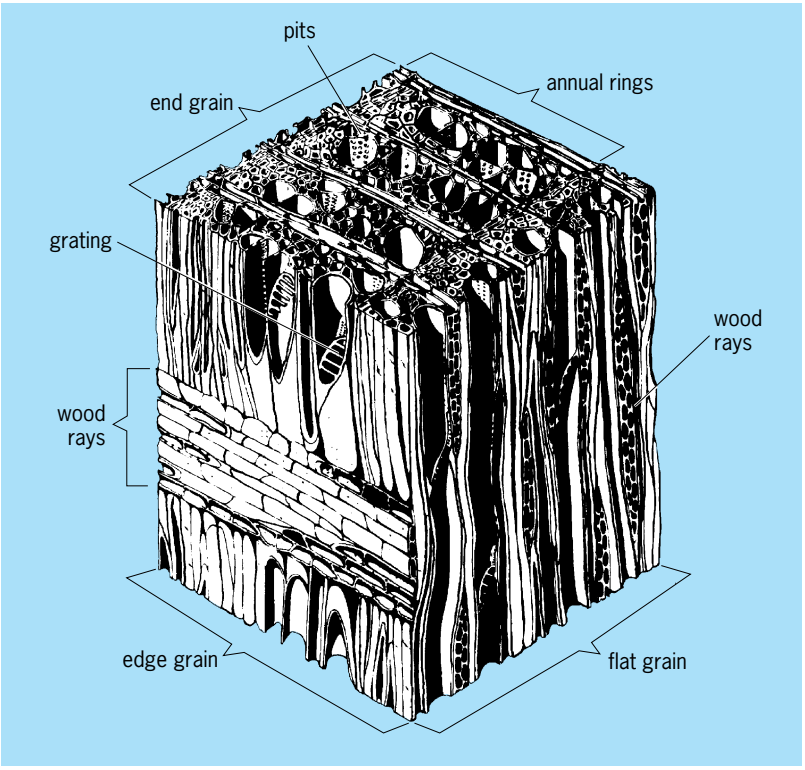


Fig. 1. Structure of a typical hardwood. (*USDA*)

tissue about which the first wood growth takes place in the newly formed twigs. *See* STEM.

The combined concentric bands of light and dark areas constitute annual growth rings. The age of a tree may be determined by counting these rings at the stump. *See* XYLEM.

In temperate climates, trees often produce distinct growth layers. These increments are called growth rings or annual rings when associated with yearly growth; many tropical trees, however, lack growth rings. These rings vary in width according to environmental conditions. Where there is visible contrast within a single growth ring, the first-formed layer is called earlywood and the remainder latewood. The earlywood cells are usually larger and the cell walls thinner than the latewood cells. With the naked eye or a hand lens, earlywood is shown to be generally lighter in color than latewood.

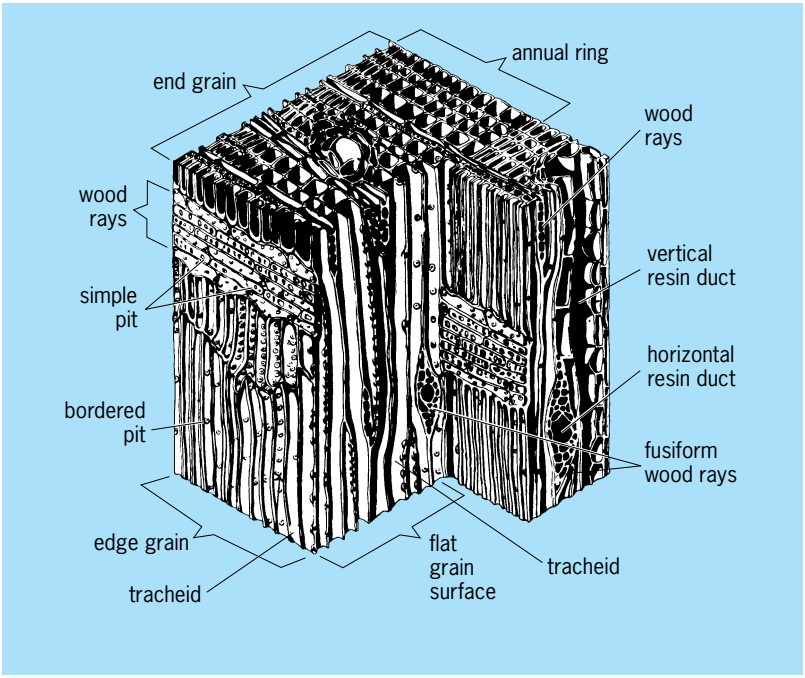


Fig. 2. Structure of a typical softwood. (*USDA*)

Because of the extreme structural variations in wood, there are many possibilities for selecting a species for a specific purpose. Some species (for example, spruce) combine light weight with relatively high stiffness and bending strength. Very heavy woods (for example, lignumvitae) are extremely hard and resistant to abrasion. A very light wood (such as balsa) has high thermal insulation value; hickory has extremely high shock resistance; mahogany has excellent dimensional stability.

Many mechanical properties of wood, such as bending strength, crushing strength, and hardness, depend upon the density of wood; the heavier woods are generally stronger. Wood density is determined largely by the relative thickness of the cell wall and the proportions of thick- and thin-walled cells present. *See* WOOD PROPERTIES.

Typical hardwood. The horizontal plane of a block of hardwood (for example, oak or maple) corresponds to a minute portion of the top surface of a stump or end surface of a log. The vertical plane corresponds to a surface cut parallel to the radius and parallel to the wood rays. The vertical plane corresponds to a surface cut at right angles to the radius and the wood rays, or tangentially within the log. In hardwoods, these three major planes along which wood may be cut are known commonly as end-grain, quarter-sawed (edge-grain) and plain-sawed (flat-grain) surfaces (**Fig. 1**).

Hardwoods have specialized structures called vessels for conducting sap upward. Vessels are a series of relatively large cells with open ends, set one above the other and continuing as open passages for long distances. In most hardwoods, the ends of the individual cells are entirely open; in others, they are separated by a grating. On the end grain, vessels appear as holes and are termed pores. The size, shape, and arrangement of pores vary considerably between species, but are relatively constant within a species.

Most smaller cells on the end grain are wood fibers which are the strength-giving elements of hardwoods. They usually have small cavities and relatively thick walls. Thin places or pits in the walls of the wood fibers and vessels allow sap to pass from one cavity to another. Wood rays are strips of short horizontal cells that extend in a radial direction. Their function is food storage and lateral conduction. Most of the rays in flat-grain surfaces are two to five cells wide, but their width and height vary in different species of hardwoods from 1 to more than 50 cells wide and from less than 100 to more than 4 in. (10 cm) in height. *See* PARENCHYMA; SECRETORY STRUCTURES (PLANT).

Typical softwood. The rectangular units that make up the end grain of softwood are sections through long vertical cells called tracheids or fibers (**Fig. 2**). Because softwoods do not contain vessel cells, the tracheids serve the dual function of transporting sap vertically and giving strength to the wood. Softwood fibers range from about 0.1 to 0.3 in. (3 to 8 mm) in length.

The wood rays store and distribute sap horizontally. Fusiform wood rays are rays with horizontal resin ducts at their centers. In the center of the end grain is a vertical resin duct. However, some softwoods, such as cedar and true fir, do not have resin ducts. The annual ring is often divided into an earlywood zone composed of thin-walled cells and a latewood zone composed of thicker-walled cells.

Sap passes from ray parenchyma cells through simple pits, unthickened portions of the cell wall, to tracheids or vice versa. Bordered pits have their margins overhung by the surrounding cell walls, but still function as passageways for sap to move from one cell to another.

Cell walls. The principal compound in mature wood cells is cellulose, a polysaccharide of repeating glucose molecules which may reach 4 μm in length. These cellulose molecules are arranged in an orderly manner into structures about 10–25 nm wide called microfibrils. This ordered arrangement in certain parts (micelles) gives the cell wall crystalline properties that can be observed in polarized light with a light microscope. The microfibrils wind together like strands in a cable to form macrofibrils that measure about 0.5 μm in width and may reach 4 μm in length. These cables are as strong as an equivalent thickness of steel.

This framework of cellulose macrofibrils is cross-linked with hemicelluloses, pectins, and lignin. Lignin, the second most abundant polymer found in plants, gives the cell wall rigidity and the substance that cements the cells together. *See* CELL WALLS (PLANT); CELLULOSE; LIGNIN; PECTIN.

Wood identification. Naked-eye field identification of unknown woods can often be made on the basis of readily visible characteristics, such as color, odor, density, or grain pattern. Observing the smoothed transverse surface with the aid of a hand lens increases the accuracy of identification, especially hardwood identification. However, for the most accurate identification, the naked eye, hand lens, and light microscope are used to examine the transverse, radial, and tangential surfaces and the various characteristics therein of the unknown wood. Wood descriptions, dichotomous keys, edge-punched cards, tables, photographs, and computer-assisted systems are also available to help in the identification process. *See* OPTICAL MICROSCOPE; PLANT ANATOMY; TREE.

Regis B. Miller

Bibliography. H. A. Core, W. A. Cote, and A. C. Day, *Wood Structure and Identification*, 2d ed., 1979; R. B. Hoadley, *Identifying Wood: Accurate Results with Simple Tools*, 1990.

Wood chemicals

Chemicals obtained from wood. The practice was carried out in the past, and continues wherever technical utility and economic conditions have combined to make it feasible. Woody plants comprise the greatest part of the organic materials produced by photosynthesis on a renewable basis, and were the precursors of the fossil coal deposits. Future shortages of the fossil hydrocarbons from which most organic chemicals are derived may result in the economic feasibility of the production of these chemicals from wood.

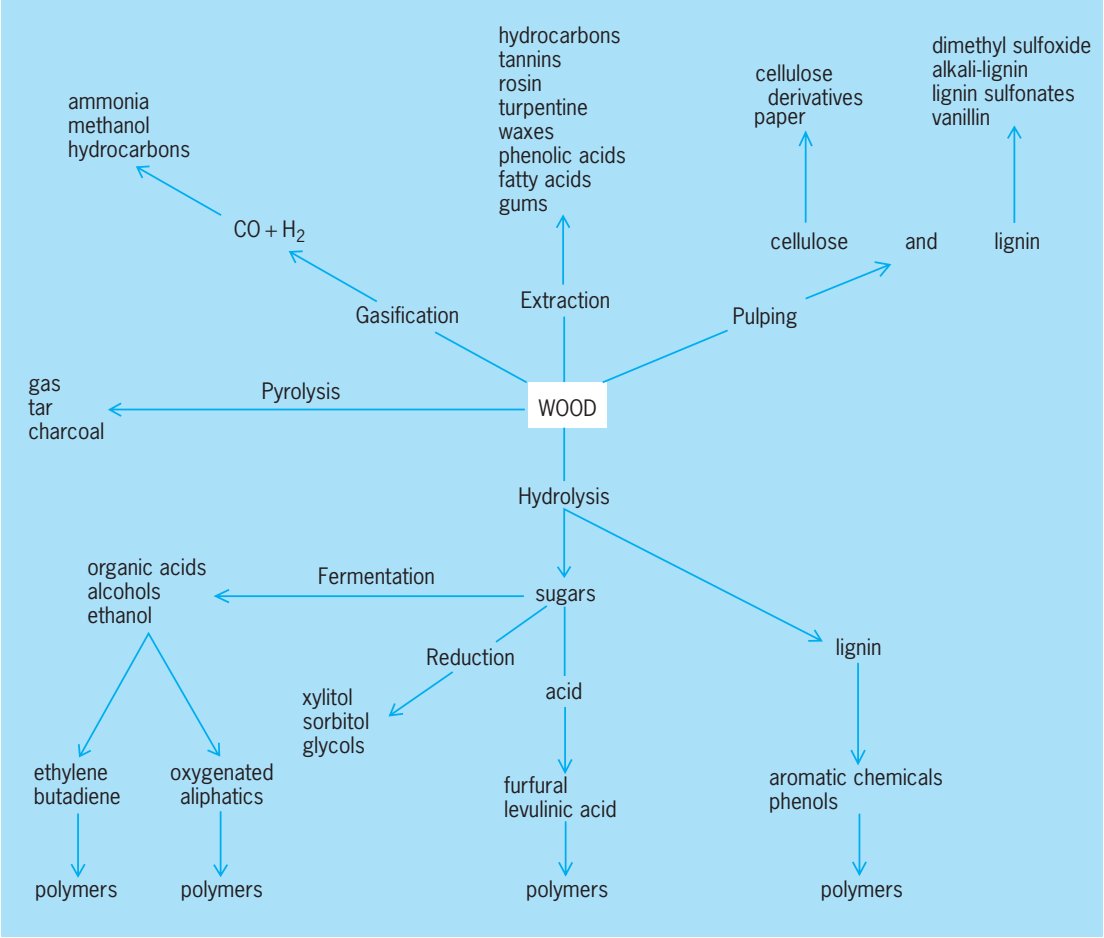
Wood is a mixture of three natural polymers—cellulose, hemicelluloses, and lignin—in an approximate abundance of 50:25:25. In addition to these polymeric cell wall components which make up the major portion of the wood, different species contain varying amounts and kinds of extraneous materials called extractives. Cellulose is a long-chain polymer of glucose which is embedded in an amorphous matrix of the hemicelluloses and lignin. Hemicelluloses are shorter or branched polymers of five- and six-carbon sugars other than glucose. Lignin is a three-dimensional polymer formed of phenylpropane units. Thus the nature of the chemicals derived from wood depends on the wood component involved. *See* CELLULOSE; HEMICELLULOSE.

Destructive distillation. In the past, destructive distillation of wood was employed to produce charcoal and was once an important industry, of which only vestiges remain. Volatile organic chemicals which were recovered from the distillate were acetic acid, acetone, and methanol. In addition, wood pyrolysis yielded tar oil fractions useful for medicinals, smoking meats, disinfectants, and weed killers. The resinous exudates from pine trees provided turpentine and rosin, while extracts of hardwoods and various barks yielded tannins useful in making leather. *See* ACETIC ACID; ACETONE; LEATHER AND FUR PROCESSING; METHANOL.

Modern processes. Chemicals derived from wood at present include bark products, cellulose, cellulose esters, cellulose ethers, charcoal, dimethyl sulfoxide, ethyl alcohol, fatty acids, furfural, hemicellulose extracts, kraft lignin, lignin sulfonates, pine oil, rayons, rosin, sugars, tall oil, turpentine, and vanillin.

Most of these are either direct products or byproducts of wood pulping, in which the lignin that cements the wood fibers together and stiffens them is dissolved away from the cellulose. High-purity chemical cellulose or dissolving pulp is the starting material for such polymeric cellulose derivatives as viscose rayon and cellophane (regenerated celluloses from the xanthate derivative in fiber or film form), cellulose esters such as the acetate and butyrate for fiber, film, and molding applications, and cellulose ethers such as carboxymethylcellulose, ethylcellulose, and hydroxyethylcellulose for use as gums.

The chemical fragments of the cell wall polymers which end up in solution after pulping can be isolated from the pulping liquors and used. These are principally lignin derivatives. Sulfonated lignins from



Chemical pathways for obtaining chemicals from wood.

the sulfite pulping process can be precipitated and used as tanning agents, adhesives, binders, dispersants, and so on. Mild alkaline oxidation of lignin sulfonates yields vanillin for flavoring and odorant applications. Alkali lignin from sulfate or kraft black liquor can be used as an extender for resins, for rubber reinforcement, and in emulsion stabilization. A volatile product from kraft black liquor is dimethyl sulfoxide, useful as a solvent. *See* VANILLA.

Sugars in spent sulfite liquor can be fermented by yeast to produce ethyl alcohol and food and fodder supplements. Oleoresinous wood components recovered from the draft pulping process include turpentine, pine oils, fatty acids, rosin, and tall oil. *See* ETHYL ALCOHOL; PINE TERPENE; ROSIN; TALL OIL.

Arabinogalactan, a hemicellulose gum extracted from larch, can be used in place of gum arabic. Bark extracts include phenolic acids from various conifers useful as extenders for synthetic resin adhesives and as binders, as well as waxes extracted from Douglas-fir bark which can be used for general wax applications. *See* WAX, ANIMAL AND VEGETABLE.

Wood hydrolysis converts the carbohydrate polymers in wood to simple sugars by chemical reaction with water in the presence of acid catalysts. Cellulose yields glucose, while softwood hemicellulose principally xylose. The hexoses can be readily fermented to ethanol. Acid treatment of hexoses yields hydroxymethylfurfural which is rapidly converted to levulinic acid, while pentoses yield furfural.

Potential chemicals. Considerable development effort has been devoted to the conversion of renewable biomass, of which wood is the major component, into the chemicals usually derived from petroleum. Processes for which technical feasibility has been demonstrated are shown in the **illustration**. Economic feasibility is influenced by fossil hydrocarbon cost and availability. *See* BIOMASS.

The wood cell wall polymers in their natural mixed state can be broken down into simpler compounds by nonselective drastic processes of pyrolysis and gasification in the same way that coal is converted into chemicals. In gasification, wood is heated at temperatures of about 1000°C (1830°F) to form a mixture of carbon monoxide and hydrogen as the major products with accompanying low yields of hydrocarbons as by-products. The carbon monoxide and hydrogen, as in coal gasification, may be converted into ammonia, methanol, or hydrocarbons. Pyrolysis or thermal degradation of wood in the absence of oxygen at lower temperatures converts the wood to charcoal, gas, tar, and oil. *See* COAL GASIFICATION; PYROLYSIS.

Selective processing of the wood components can

retain more carbon-carbon bonds and provide substantial yields of a wide variety of chemicals. Hydrolysis of cellulose to glucose by high-temperature dilute acids, low-temperature strong acids, or enzymes is the first step in its utilization for chemicals. Ethanol from glucose fermentation may be used as an industrial chemical, as a fuel for internal combustion engines, or as an intermediate for the production of other chemicals.

Ethanol may be dehydrated to ethylene, which is the building block for many other organic intermediates and polymers. By oxidation of ethanol to acetaldehyde, an intermediate for further production of acetic acid and anhydride, acrylonitrile, vinyl acetate, and butadiene may be obtained.

Fermentation of glucose may alternatively be directed to lactic acid, acetone, butanol, isopropanol, glycerin, and so on. Glucose hydrogenation can yield glycols as well as sorbitol. *See* FERMENTATION.

Hydrolysis of hemicelluloses yields mannose, which can be processed along with glucose, and xylose, which can be converted to furfural or xylitol. The aromatic lignin component of wood can be converted into phenols and other aromatic chemicals such as benzene under various hydrogenation conditions followed by dealkylation and dehydroxylation. *See* ORGANIC SYNTHESIS; WOOD PRODUCTS.

Irving S. Goldstein

Bibliography. I. S. Goldstein, *Organic Chemicals from Biomass*, 1981; P. Hakkila, *Utilization of Residual Forest Biomass*, 1989; E. J. Soltes, *Wood and Agricultural Residues: Research on Use for Feed*, *Fuels*, and *Chemicals*, 1983; D. L. Wise, *Organic Chemicals from Biomass*, 1983.

Wood composites

Wood-based composites and a combination of two or more elements held together by a matrix. By this definition, what we call "solid wood" is a composite. Solid wood is a three-dimensional composite composed of cellulose, hemicelluloses, and lignin polymers with smaller amounts of inorganics and extractives held together in a lignin matrix. Among the advantages of developing wood composites from solid wood are the ability to: (1) use trees or tree parts that are too small for lumber or veneer production (includes thinning, limbs, brush, and other forest biomass), (2) use underutilized wood species, (3) use mixed wood species, (4) remove defects, (5) create more uniform materials, (6) develop materials that are stronger than the original solid wood, (7) make materials of different shapes, and (8) produce both large-volume low-value and low-volume value-added materials. A beneficial by-product is the removal of biomass from overcrowded forests, especially in the western United States, to improve forest health and to reduce the fuel load that increases the threat of catastrophic fires.

Historically, wood was used only in its solid form as large timbers, or lumber. As the availability of large-diameter trees decreased and the price increased, the wood industry looked to replace large timber products and solid lumber with reconstituted wood products made using smaller diameter trees, and manufacturing, saw and pulp mill wastes.

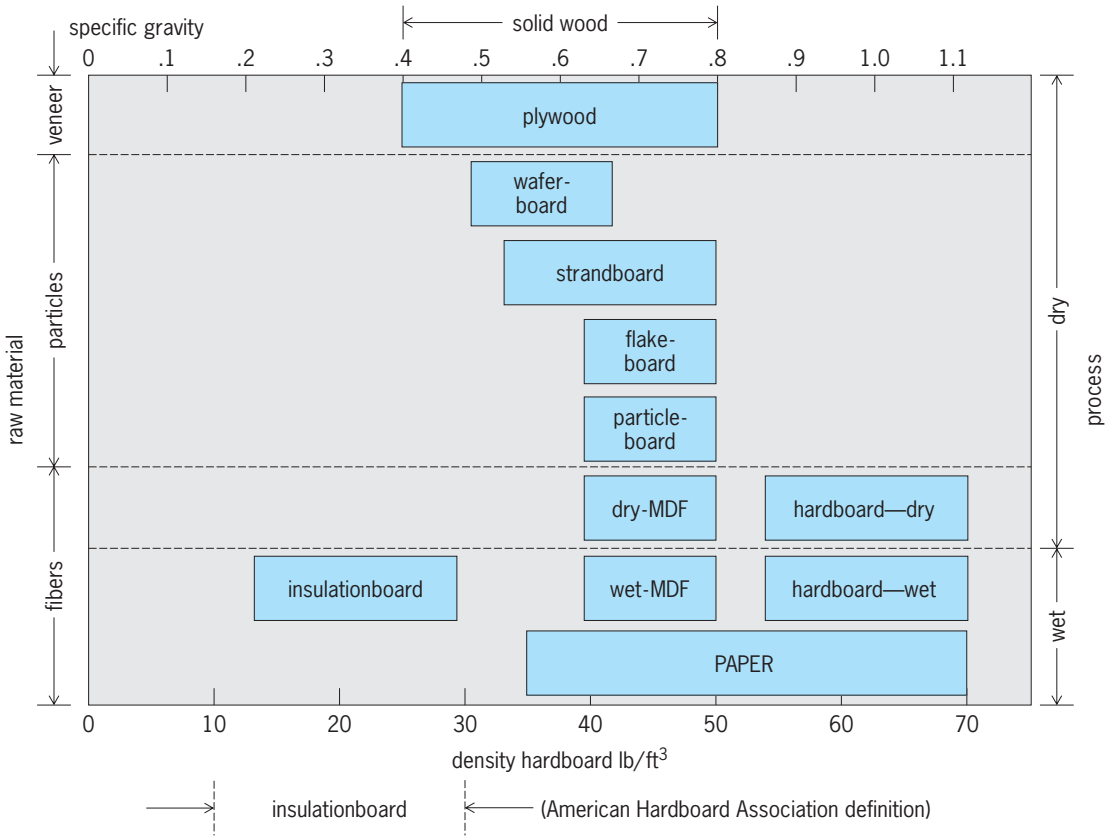
Wood elements. Solid wood can be broken down into smaller and smaller elements, that is, veneers, flakes, strands, chips, flakes, particles, fiber, and cellulose. As the element size becomes smaller, uniformity increases and composites made from these smaller elements become more like a true material, that is, consistent, uniform, continuous, predictable, and reproducible. As the size of the composite element gets smaller, it is also possible to either remove defects (such as knots, cracks, and checks) or redistribute them to reduce their effect on product properties. Size reduction, however, requires energy and may increase water consumption. Adhesion and the development of the composite matrix is also a critical issue.

Adhesives. The basic adhesives most commonly used for wood composites are formaldehyde, urea, melamine, phenol, resorcinol, and isocyanates. Despite the apparent simplicity in terms of families of chemicals, the formulations are highly complex mixtures of chemicals and additives depending on the specific application. Wood composites used outdoors are usually produced using phenol, resorcinol, or isocyanate adhesives.

Types of composites. Wood composites can be classified in several ways: by density, element type, process, or application. Classification by specific gravity includes high-, medium-, and low-density composites (see **illustration**). Classification by element type includes strandboard, waferboard, chipboard, flakeboard, particleboard, and fiberboard. Classification by process includes wet or dry production processes. Classification by application would include products such as insulation board, underlayment composites, and structural and core composites.

Strandboard. Structural composite lumber (SCL) can be made using strands of small-diameter wood glued parallel to the length. SCL products include oriented strand lumber (OSL) and parallel strand lumber (PSL). Laminated strand lumber (LSL), oriented strand board (OSB), and OSL are produced using different lengths and sizes of strands. LSL uses strands that are about 0.3 m in length while OSB is produced from shorter strands. PSL is made from strands that are 3 mm thick, approximately 19 mm wide, and 0.6 m in length. Usually Douglas-fir, southern pines, western hemlock, and yellow poplar are used but other species are also used. The major adhesives used to produce SCL products are phenol-formaldehyde or isocyanates. All of these SCL products are used as replacements for solid wood and have a specific gravity of 0.5 to 0.8.

Plywood. The modern plywood industry began around 1910 but the furniture industry had used veneers over solid wood for several hundred years before that. Plywood is usually made using veneers placed at right angles to each other to increase dimensional stability and add strength.



Wood composites classified by density, element type, process, and application.

Plywood is mainly used as underlayment in roofs and flooring.

Waferboard and flakeboard. The wafer board and flakeboard industries started in the 1960s. The specific gravity of wafer and flakeboard is usually between 0.6 and 0.8 and made using a waterproof adhesive such as phenol formaldehyde or an isocyanates. Flakes used in flakeboard production vary in size from 1 to 2 cm wide, 2 to 3 cm in length, and 2–5 mm in thickness. Wafers are almost as wide as they are long, while flakes are much longer than they are wide. Wafers are also thicker than flakes. These are used as the structural skin over wall and floor joists. *See* PLYWOOD.

Particleboard. The particleboard industry started in the 1940s. Particleboard is produced using particles that vary in length from about 2 mm to 1 cm. The specific gravity of particleboard is between 0.6 and 0.8 and is usually produced from softwoods such as Douglas-fir, southern pines, or other low-value wood sources. Today, particleboard is used in furniture using metal or nylon screw-together joints.

Hardboards. The hardboard industry started around 1950 and is produced mainly by a dry process with a specific gravity between 0.8 and 1.1. Dry formed hardboard is made using a fiber made by mechanical separation of the fibers from the wood. The medium-density fiberboard (MDF) industries started in the early 1960s and are usually made from fiber processed by aqueous high-temperature refining.

Three-dimensional shapes. Strand type boards can be formed into three-dimensional shapes but, in general, all of the products described above are produced in flat sheets and used in two-dimensional designs. Three-dimensional composites can also be made using mainly particles and fibers and the final shape can be formed in the hot press during manufacturing. *See* WOOD ENGINEERING DESIGN; WOOD PRODUCTS; WOOD PROPERTIES.

Roger M. Rowell

Bibliography. *Panel Trends*, 2002, Composite Panel Association. Gaithersburg, MD. 2002; *Regional Production & Market Outlook for Structural Panels and Engineered Wood Products, 2002–2007*, APA—Engineered Wood Association, Tacoma, WA, 2002; *Wood Product Demand and the Environment: Proceedings of an International Conference Sponsored by the Forest Products Research Society*, Madison, WI, 1992; *Wood Products for Engineered Structures: Issues Affecting Growth and Acceptance of Engineered Wood Products*, Forest Products Research Society, Madison, WI, 1993.

Wood degradation

Decay of the components of wood. Despite its highly integrated matrix of cellulose, hemicellulose, and lignin, which gives wood superior strength properties and a marked resistance to chemical and microbial attack, a variety of organisms and processes are capable of degrading wood. The decay process is a continuum, often involving a number of organisms over many years. Wood degrading agents are both biotic and abiotic, and include heat, strong acids or

bases, organic chemicals, mechanical wear, and sunlight (uv degradation).

Abiotic degradation. Heat degrades both cellulose and hemicellulose, reducing strength and causing the wood to darken. At temperatures above 451°F (219°C), combustion occurs.

Because wood is resistant to many chemicals, it has long been used in cooling towers. Strong acids eventually degrade the carbohydrate portion of wood, reducing its strength. Strong bases attack the lignin, leaving the wood appearance bleached and white; this effect forms the basis for the Kraft pulping of wood chips for paper production. Other chemicals, such as concentrated organics or salt solutions, can also disrupt the lignocellulosic matrix, reducing material properties of the wood. *See* PAPER.

Sunlight, primarily through the action of ultraviolet light, also degrades wood through the creation of free radicals which then degrade the wood polymers. Ultraviolet degradation extends only a few cell layers from the surface, but can become severe when water or wind continually removes weakened wood, exposing nondegraded wood beneath to ultraviolet light. In most instances, however, ultraviolet degradation causes the wood to change to a gray to brown color which is highly valued in some markets.

Mechanical wear of wood can occur in a variety of environments. For example, floating docks in marinas may shift up and down along a wood pile, continually abrading the wood and reducing the effective pile circumference. Wood railway ties are subjected to repeated shock loading as rail cars pass over. Ties in mainline track often fail from mechanical effects, long before other agents of deterioration can cause damage.

Abiotic degradation is often difficult to distinguish from biotic damage. However, the absence of biological organisms plus increases in pH or the location of the damage can provide important clues concerning the nature of the damage.

Biotic degradation. Biotic damage can occur from a variety of agents, including bacteria, fungi, insects, marine borers, and birds and animals. Birds and animals generally cause mechanical damage in isolated instances.

Biotic requirements. All biotic agents have four basic requirements: adequate temperature (32–104°F or 0–40°C) with most optima between 77–90°F (25–32°C), oxygen (or other suitable terminal election acceptor), water, and a food source. Temperature ranges for growth of decay organisms vary widely, and temperature is generally not controllable as a means for limiting the activities of wood-degrading agents.

Although oxygen is required for activity of most wood-degrading organisms, some bacteria are capable of anaerobic growth. Furthermore, many fungi are capable of growing at very low oxygen tensions or can become dormant under low oxygen conditions and then resume growth once oxygen levels improve. Many lumber mills create low oxygen conditions by submerging logs in ponds to limit biologic attack until the wood can be sawn and dried. Prolonged submersion, however, permits bacterial degradation of wood pit membranes, increasing wood permeability and causing logs to sink to the bottom of the pond.

Water is a critical element for biotic decay agents: it serves as reactant in degradative reactions, a medium for diffusion of enzymes into wood and degradative products back to the organism, and a wood swelling agent. Wood is considered most decay susceptible when the moisture content exceeds 30% by weight. At this level, called the fiber saturation point; free water is present in the wood. Thus, keeping wood dry is the primary method for protecting against decay, although some organisms have adapted to survive in wood at lower moisture levels.

Bacteria. Bacteria are primitive, single-celled organisms that generally are not major degraders of wood products, but they can damage pit membranes, thereby increasing permeability, and some are capable of cell wall degradation. Bacteria appear to be most important in submerged environments, highly fertilized agricultural soils, and other regimes where growth by more common fungal agents is limited. The role of these agents and their interactions with other wood-degrading organisms remain poorly defined.

Fungal organisms. Fungi are among the most important wood-degrading organisms because they play an important role in terrestrial carbon cycling. Wood-degrading fungi can be classified as molds, stainers, soft rotters, brown rotters, and white rotters on the basis of the attack patterns. Molds, stainers, and soft rotters are members of the ascomycetes and the deuteromycetes (Fungi Imperfecti). Molds primarily colonize the surface of freshly sawn wood, attacking readily available carbohydrates in the ray cells. These fungi cause cosmetic damage to the wood due to the production of pigmented spores (which can be brushed from the wood surface), and they can increase wood permeability, leading to uneven finishing of wood. *See* ASCOMYCOTA; DEUTEROMYCOTINA.

Stain fungi cause more serious damage since they are cellulolytic and can remove large portions of the ray cells. The hyphae of stain fungi are dark pigmented, giving the wood a bluish color. Stain fungi cause some loss in toughness and increase permeability, but have little effect on most properties. Both molds and stain fungi can be controlled by rapid drying after sawing or by application of prophylactic fungicides shortly after sawing.

Soft rot fungi are capable of significant degradation of wood carbohydrates in specialized environments, particularly where wood is wet or exposed to exogenous nutrients. Soft rot attack is often concentrated near the wood surface, and the decay quickly reduces strength. These fungi also exhibit significant preservative tolerance and are an important factor in degradation of some wood species.

Brown rot fungi are basidiomycetes and use wood carbohydrates in a manner which reduces strength rapidly at the early stages of attack. Brown rotters can remove up to 70% of the wood mass, leaving a cracked brown degraded mass of modified lignin. Brown rot fungi are most prevalent in coniferous

wood species and are associated with dramatic drops in wood strength at very early stages of decay. As a result, early detection of these fungi is an important aspect of a wood inspection program.

White rot fungi are also basidiomycetes, but these species can utilize all wood components, removing up to 97% of the wood. White rot fungi are most prevalent on hardwoods, possibly owing to a reduced lignin content in these species. The ability of white rot fungi to degrade complex ring structures has been exploited for degradation of pesticides and pretreating chips for manufacturing paper. *See* BASIDIOMYCOTA; FUNGI; PESTICIDE.

Insects. A number of insects have evolved to attack wood, including termites (Isoptera), beetles (Coleoptera), and bees and ants (Hymenoptera). Termites are social insects with highly organized worker, soldier, and reproductive castes. They are the most important wood-degrading insects in most environments, and their activity causes severe economic losses. Termite damage is generally difficult to detect because the workers do not forage outside. As a result, most damage is detected when reproductives swarm from the nest to start new colonies or when wood is so badly damaged that it fails. Most termites require wet wood, so breaking wood-soil contact and preventing wetting are preventive measures. Drenching the soil around a structure with a termiticide to develop a barrier is also commonly used. *See* COLEOPTERA; HYMENOPTERA; ISOPTERA.

A variety of beetles attack freshly fallen trees with the bark attached, but most do not attack finished products, although their damage may be found in the material. Of the wood-attacking beetles, the powder post beetles (comprising families Anobiidae, Bostrichiidae, and Lyctidae) are among the most important. Powder post beetles often cause problems in structures which are subject to infrequent inspections, and are a problem in museums. Damage is usually detected by the presence of piles of finely powdered frass below a wood member or by the presence of small exit holes on the wood surface. Powder post beetle infestations are often localized, and control is usually effected by fumigation with methyl bromide or sulfuryl fluoride.

Among the Hymenoptera which attack wood are the carpenter bees (Anthophoridae) and carpenter ants (Formicidae). Neither of these insects uses the wood as a food source. Carpenter bees excavate long galleries along the grain and provision cells along the tunnel with captured insects. Carpenter ants are social insects with a colony structure similar to the termites. They are commonly found in moist wood, and their damage can be substantial. Insecticides sprayed outside the nest and transported into the nest by workers are most effective since they have a higher probability of killing the queen.

Marine borers. In saline environments, marine borers can cause significant wood losses. While marine borers are often considered to be detrimental, these organisms are important recyclers of coarse woody debris in marine environments and their droppings provide a nutrient source to deep-ocean communities. Three groups of marine borers [shipworms, pholads, and gribbles (*Limnoria*)] cause most wood damage in these areas. Shipworms and pholads are clamlike mollusks which begin life as free-swimming larvae which then settle on the wood surface. Pholads continue to remain in their shells and gradually create pear-shaped cavities near the wood surface, from which they filter-feed in the surrounding water. Shipworms use two small shells at the top of the head to rasp a wood cavity and penetrate deeper into the wood, where they can grow to lengths of up to 5 ft (1.5 m) in 2–3 years. Shipworms are difficult to detect, and their internal damage can lead to wood failure in as little as 2 years. Both shipworms and pholads can be excluded from wood by treatment with creosote, while shipworms can also be controlled with a number of inorganic arsenicals. *See* SHIPWORM.

Limnoria are mobile crustaceans which tunnel into wood near the surface, but are not believed to use wood as a food source. The tunnels weaken the wood, which gradually erodes because of wave action. The freshly exposed wood is also attacked, creating a cycle of attack and erosion which eventually results in wood failure. Damage to wood by *Limnoria* is most often prevented by treating the wood with creosote or inorganic arsenicals. In areas where pholads and *Limnoria* attack, dual creosote and inorganic arsenical treatments are required. *See* BORING BIVALVES.

Wood protection. Protecting wood from degradation can take a number of forms. By far the simplest method is to employ designs which limit wood exposure to moisture, and the effectiveness of these methods can be seen in many centuries-old wood buildings in Asia and Europe. In some cases, however, water exclusion is not possible and alternative methods must be employed. The simplest of these methods is the use of heartwood from naturally durable species. Sapwood of any species has no natural durability, but some wood species produce heartwood which is resistant to biological attack. Decay- or insect-resistant species include redwood (*Sequoia sempervirens*), western red cedar (*Thuja plicata*), and ekki (*Lophira alata*), while marine-borer-resistant heartwoods include greenheart (*Ocotea rodiaei*) and ekki. Most marine-borer-resistant woods contain high levels of silica which discourages marine borer attack, while species resistant to terrestrial decay agents often contain toxic phenolics.

Wood can also be protected from degradation by spraying, dipping, soaking, or pressure treatment with preservatives. Protecting wood from mold and stain fungi is most often accomplished by dipping or spraying the wood with a fungicide shortly after sawing. These treatments generally provide up to 6 months of protection to the wood surface. *See* WOOD PROPERTIES.

Jeffrey J. Morrell

Bibliography. R. A. Eaton and M. D. Hale, *Wood: Decay, Pests and Protection*, 1993; A. S. Panshin and C. de Zeeuw, *Textbook of Wood Technology*, 4th ed., 1980; R. A. Zabel and J. J. Morrell, *Wood Microbiology: Decay and Its Prevention*, 1997.

Wood engineering design

The process of creating products, components, and structural systems with wood and wood-based materials. Wood engineering design applies concepts of engineering in the design of systems and products that must carry loads and perform in a safe and serviceable fashion. Common examples include structural systems such as buildings or electric power transmission structures, components such as trusses or prefabricated stressed-skin panels, and products such as furniture or pallets and containers. The design process considers the shape, size, physical and mechanical properties of the materials, type and size of the connections, and the type of system response needed to resist both stationary and moving (dynamic) loads, and function satisfactorily in the end-use environment. *See* ENGINEERING DESIGN; STRUCTURAL DESIGN.

Wood is used in both light frame or heavy timber structures. Light frame structures consist of many relatively small wood elements such as lumber covered with a sheathing material such as plywood. The lumber and sheathing are connected to act together as a system in resisting loads; an example is a residential house wood floor system where the plywood is nailed to lumber bending members or joists. In this system, no one joist is heavily loaded because the sheathing spreads the load out over many joists. Service factors such as deflection or vibration often govern the design of floor systems rather than strength. Light frame systems are often designed as diaphragms or shear walls to resist lateral forces resulting from wind or earthquake. *See* FLOOR CONSTRUCTION.

In heavy timber construction, such as bridges or industrial buildings, there is less reliance on system action and, in general, large beams or columns carry more load transmitted through decking or panel assemblies. Strength, rather than deflection, often governs the selection of member size and connections. There are many variants of wood construction using poles, wood shells, folded plates, prefabricated panels, logs, and combinations with other materials.

Solid wood. Structural dimension lumber is the most commonly used solid wood engineering material, particularly for light frame structures as well as for residential roof and floor trusses. Nonstructural grades of lumber assure satisfactory performance in furniture frames, pallets and containers, truck flooring, and a myriad or other products. The strength, stiffness, and other properties of structural lumber are determined by grading each piece. A visual stress rating is determined by looking at each piece and assigning it to a strength category based on the observed defects and growth characteristics. Machine stress rating classifies lumber by using a machine to nondestructively test each piece. This method is supplemented by additional human control for characteristics which cannot be physically tested. The stress rating process accounts for the different inherent properties of clear, defect-free wood and the influence of growth characteristics such as knots or sloping grain.

Mechanical properties. The use of specific wood types for engineering and other purposes is related to the wood's mechanical properties. Growth and manufacturing characteristics such as knots, grain deviation, and wane reduce strength of structural lumber, and increase its variability. Allowable safe design strength and stiffness for lumber depends on the lumber grade. The tension perpendicular to grain and the shear strength for wood are relatively low and dictate that wood products should not be used where these stresses may dominate. For example, short deep wood beams with a span/depth ratio less than 12 may not be efficient. Also, high loads should not be fastened to the lower side of wood beams because of cross-grain tension.

Dry wood is generally ductile and flexible with excellent energy absorption characteristics except in tension perpendicular to grain and shear where failure may be preceded by little warning. The strength of wood members depends on the duration of an applied load under some conditions. In general, wood has greater resistance to short-term loads than to long-term loads. Limits on the deflection of beams and the buckling capacity of long columns may govern the design of these elements because of moderate stiffness. Wood columns that are short or fully braced are usually economic if buckling is avoided. An example is the use of wood in pilings, piers, and utility poles.

High temperatures, above 150°F (65°C), can cause permanent strength and stiffness reductions in wood under prolonged exposure. Increasing wood moisture content can also reduce some properties and require that wood be treated to prevent biodeterioration. *See* FOUNDATIONS; WOOD ANATOMY; WOOD PROCESSING; WOOD PROPERTIES.

Engineered wood composites. Wood composites are products composed of wood elements that have been glued together to make a different, more useful or more economical product than solid sawn wood. Plywood is a common example of a wood-composite sheathing panel product where layers of veneer are glued together. Plywood is used as a sheathing material in light frame wood buildings, wood pallets, and containers to distribute the applied forces to beams of lumber or other materials. Shear strength, bending strength, and stiffness are the most important properties for these applications and may be engineered into the panel by adjusting the species and quality of veneer used in the manufacture.

Glue-laminated timber (glulam) is a similar example where layers of lumber are glued together to form larger, longer, and stronger structural elements than can be obtained with solid sawn lumber alone. Glulam beams can reach lengths over 100 ft (30 m), with cross sections exceeding 12 × 36 in. (300 × 900 mm). Curved glulam beams and arches are manufactured to maintain the intrinsic strength of wood by keeping the grain direction running along the length of the element.

Laminated veneer lumber is manufactured by

gluing together layers of dry veneer into a large panel with the grain direction of all layers oriented in the same direction. The panel is then ripped and crosscut to form a lumberlike product. The depth of the lumber is determined by sawing, and the width is determined by the number of veneer layers. Laminated veneer lumber generally has greater strength than solid sawn lumber because the defects and growth characteristics are distributed through-out the element rather than concentrated in a knot or some other characteristic. In addition, this type of lumber can be made from one or many species and cut into sizes not obtainable with solid sawn lumber.

The strength of laminated veneer lumber is determined by testing and is ensured by continuous quality control over the manufacturing process. Laminated veneer lumber is used for high-strength applications where straightness is important, such as truss chords, composite I beams, garage and window headers, scaffold planking, and furniture frames. Laminated veneer lumber can also be made with some curvature by pressing the glued veneer in a curved form before the adhesive cures.

Oriented strand board. This composite panel sheathing product is made by gluing thin flakes or compressed strands under pressure and heat. The flakes or strands are lined up and arranged in three to five layers which are oriented at right angles to one another. The center layer may have strands in a random orientation. This orientation by layers gives more uniform strength and stiffness to the panel in the same manner as alternating veneer grain direction does in plywood. Oriented strand board is used the same way as plywood and has similar strength and stiffness properties. In addition, oriented strand board manufacturers can use irregularly shaped logs and low-density species to make the flakes and strands. The properties are ensured by test and continuous quality control of the manufacturing process.

Composite I beams. Usually made of several wood-based materials, composite I beams are used as lumber substitutes for bending applications. This product looks like an I beam with the outer flanges designed to carry the bending moment through axial compression and tension. The center of the I or web is designed to resist the beam shear stresses. Laminated veneer lumber or solid-sawn lumber is used as the flange material which is glued to a web made of plywood or oriented strand board. Composite I beams are used in either light frame or heavy timber structures where a very uniform economical product is required. Special care must be taken to ensure the lateral stability and bearing capability of this product because of the relatively thin web.

Connections. A major advantage of wood and wood-based engineering materials is the ease with which they can be connected. Nails, staples, and screws are field installed in light frame structures by hand or with simple power equipment. This creates adequate connections that have high energy absorption and ductility. Individual connections, however, are generally not engineered with light frame construction. Fastening schedules, which specify number, size, and type of fastener for a specific application, have been developed by product manufacturers and building code officials.

The development of modern wood structural systems depends on efficient, economical connections systems to transfer forces. Prefabricated steel connectors installed with nails or bolts, such as the joist hanger, tie-downs, and anchors, have enabled engineered wood construction to meet increasing safety and serviceability demands in high-wind and earthquake areas. The punched-tooth metal connector plate has enabled the development of the prefabricated wood truss which is prevalent in both roof and floor systems. The metal connector plated truss is the most highly engineered component in the light frame structural system, and each truss connection is individually engineered.

Adhesive connections are efficient means of force transfer but require a high level of technology and quality control. They are commonly used in glulam where finger jointing provides a means of developing any length of lumber. The webs are glued to the flanges in composite I beams. Plywood is glued to the lumber ribs in stressed-skin panels to develop the full system action of the panel. However, these adhesive connections are made in a factory under controlled conditions. Although there are exceptions, the most common field use is with elastomeric adhesive connections to enhance the stiffness of the nail connection between sheathing and solid wood, thus improving the overall service performance of floor systems.

Bolts, shear plates, split rings, and a myriad of special connectors are used with glulam and heavy timber construction. Traditional connections such as mortise and tenon joints are found in timber frame construction. In general, heavy timber connections are more expensive and more localized than with light frame construction. Each connection is designed independently, and strength is the most important criterion. There are two general design philosophies: the connection should always be stronger than the more costly connected members whose failure could be catastrophic to the structure; and connections should be designed to control failure, avoid catastrophe, and allow for repair without the need for member replacement. Either philosophy, or both, can be used depending on the design objectives, the type of expected loads, and the opportunity. *See* WOOD PRODUCTS. Thomas E. McLain

Bibliography. J. Bodig and B. A. Jayne, *Mechanics of Wood and Wood Composites*, 1993; K. F. Faherty and T. G. Williamson, *Wood Engineering and Construction Handbook*, 3d ed., 1998; Forest Products Laboratory Staff, *Wood Engineering Handbook*, 2d ed., 1990; A. D. Freas, R. C. Moody, and L. A. Soltis (eds.), *Wood: Engineering Design Concepts*, 1983.

Wood processing

Peeling, slicing, sawing, and chemically altering hardwoods and softwoods to form finished products such as boards or veneer; particles or chips for making

paper, particle, or fiber products; and fuel. *See* PAPER; VENEER.

Most logs are converted to boards in a sawmill that consists of a large circular or band saw, a carriage that holds the log and moves past the saw, and small circular saws that remove excess bark and defects from the edges and ends of the boards. One method is to saw the log to boards with a single pass through several saw blades mounted on a single shaft (a gang saw). Sometimes, the outside of the log is converted to boards or chips until a rectangular center or cant remains. The cant is then processed to boards with a gang saw.

Drying. A high percentage of the weight of freshly cut or green wood is water. The cells of green wood contain free water in the cell cavities and bound water in the cell walls. When all the free water has been extracted and before any of the bound water has been removed, the wood is said to be at the fiber saturation point. This is a moisture content of about 28% of the ovendry weight of the wood. Above the fiber saturation point, wood does not change dimension with a change in moisture content.

As the moisture content falls below the fiber saturation point, the bound water leaves the cell walls and the wood shrinks. But it does not shrink uniformly because wood is anisotropic; that is, it does not have the same properties in all directions. In drying a cross section of a log, wood shrinks very little longitudinally (the direction that is parallel to the long axis of the log). Shrinkage around the outside of the log (tangential) is about twice that from the center of the log to the bark (radial) and can be as much as 10% of the green dimension. A dry piece of wood does not have the same dimensions and often not the same shape as when green. During the drying process, differential shrinkage can cause internal stresses in the wood. If not controlled, this can result in defects such as cracks, splits, and warp.

Below the fiber saturation point, wood takes on and gives off water molecules depending on the relative humidity of the air around it and swells and shrinks accordingly. The drying process is complicated because during the initial stages, when the outside of a board drops below the fiber saturation point and the inside remains above it, the outside shell is restrained from shrinking. The outside shell will also set and become larger than if the shell were unrestrained. If drying is too rapid during this period, surface checks and splits can occur. As drying continues, the outer shell of the board eventually restrains the inner core from shrinking. If drying in this latter stage is too rapid, internal checks called honeycombs can occur. *See* WOOD PROPERTIES.

A dry kiln is used to control the rate of moisture loss. Essentially, a dry kiln is a room in which a controlled flow of air at a specified temperature and relative humidity is passed through layers of carefully stacked lumber. A method of venting moisture-laden air outside the kiln is necessary. Kiln schedules that safely remove moisture have been developed for most commercially important species. For easily dried species (most softwoods and some low-density hardwoods), schedules are based on time at specific temperatures.

At the end of the drying process, lumber can contain drying stresses. Casehardening occurs when the center of the board is subjected to tension and the outside of the board is being compressed. For boards that are to be processed further, such as to furniture or cabinet parts, it is imperative that these stresses be relieved. The final steps in the kiln-drying process are called stress equalizing and conditioning.

The final moisture content at a given temperature and relative humidity is called the equilibrium moisture content. In the United States, the equilibrium moisture content usually is about 6% in heated buildings in the winter and about 12% in the summer. Therefore, products for interior use should be manufactured at moisture contents of about 7–8% to avoid in-service product problems. *See* WOOD PROPERTIES.

Machining. Wood is machined to bring it to a specific size and shape for fastening, gluing, or finishing. With the exception of lasers, which have a limited application at this time, all machining is based on a sharpened wedge that is used to sever wood fibers. Tools for sawing, boring holes, planing, and shaping, as well as the particles in sandpaper, use some version of the sharpened wedge.

The process of passing a single blade over the workpiece as with a hand plane is called orthogonal cutting. There are three tool angles that affect machining quality (**Fig. 1**). The cutting or rake angle is the angle from a line perpendicular to the workpiece to the face of the tool. The sharpness angle is measured from the face to the back of the tool. The clearance angle is from the back of the tool to the new surface of the workpiece. In general, the cutting angle determines the type of wood chip; the sharpness angle affects tool life; and the clearance angle must be sufficient to prevent rubbing of the new surface with the back of the blade.

There are three basic types of wood chips formed by knife machining: Large cutting angles cause splitting ahead of the blade (type I chip; **Fig. 2**). For each

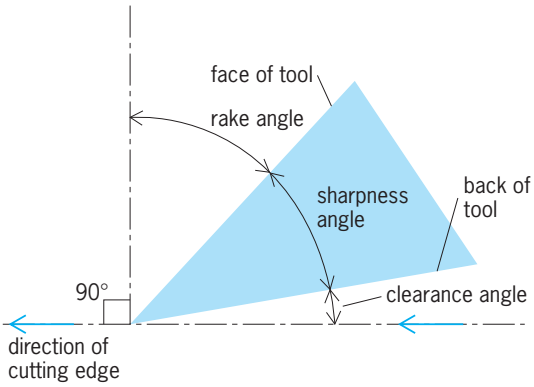


Fig. 1. The geometry of a cutting tool is described by the rake angle, measured from a line perpendicular to the direction of travel to the tool face; the sharpness angle, measured between the face and back of the tool; and the clearance angle, measured between the back of the knife and its direction of travel. (*After R. B. Hoadley, Understanding Wood, Taunton Press, 1981*)

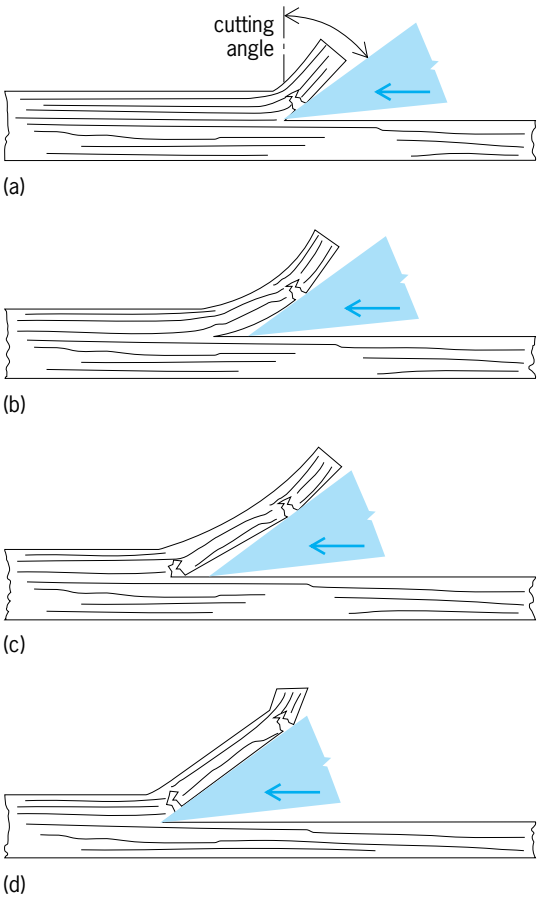


Fig. 2. Cutting action in planing wood. (*a*) Cut begins. (*b*) Chip (type I) bends, as it slides up the knife, and the wood fails ahead of the edge because of tension perpendicular to the grain. (*c*) Chip breaks. (*d*) Next segment of the cut begins. (*After R. B. Hoadley, Understanding Wood, Taunton Press, 1981*)

species and grain angle, there is a range of cutting angles that will cause a continuous chip or planer shaving and generate a smooth surface (type II chip). Small cutting angles cause the knife to plow through the wood and form a type III chip.

Small sharpness angles with large cutting and clearance angles offer the best of all possibilities. However, because relatively little tool material is present, tool dulling is relatively rapid. As the tool cutting edge rounds with wear, the cutting and clearance angles are changed at the point of fiber severance. In effect, a cutting edge that originally made a type II chip can dull so that the cutting angle becomes small and makes a type III chip.

When several blades are mounted in a rotating cutterhead and the workpiece is fed past, the process is called peripheral milling. It is difficult to manually make all cutting edges protrude equally. A process called jointing grinds a small amount from the back of the protruding cutting edge. If the protruding edge is excessive, a negative clearance angle will be present on a sharp knife and possibly cause surface defects (raised or loosened grain).

The most important factor affecting surface smoothness in knife planing is whether the knife is machining with or against the grain (**Fig. 3**). It is rare that the grain will be parallel to the surface for the entire length of a long, clear piece. In addition, the grain in and around knots often poses a problem for straight knife planers as torn or chipped grain can result. Abrasive planers that use belts of sandpaper with large grits to rough-plane are used to prevent tearing and chipping. *See* WOOD ENGINEERING DESIGN.

Veneer cutting. Veneer cutting is considered a special case of orthogonal machining in which the type II chip is the objective rather than a smooth surface on the core. Veneer is made by peeling (rotary cutting) or slicing and can be cut in thicknesses from $^1/_{100}$ to $^1/_4$ in. In rotary cutting of veneer, a log is chucked on both ends and turned while a knife advances against it. The rate of advance and the speed of log rotation determine the thickness of the veneer. For slicing, a block of wood is held against a frame that moves up and down diagonally. The slicing knife is advanced between cuts to set the thickness of the veneer.

The side of the veneer in contact with the knife can crack or check severely as it moves sharply away from the flitch or core, particularly when the veneer is thick. The loose side (with lathe or knife checks) and the tight side (without checks) can be identified by flexing the veneer. In flexing, the checks in the loose side open slightly, allowing the veneer to bend easily while the tight side remains fairly stiff. Paint or other finishes may fail if the loose side is mistakenly used as the exposed face in plywood or inlaid veneer products. To reduce or prevent lathe checks, a nose or pressure bar is set next to the knife so that the wood is compressed slightly just before it is cut.

Composite-type products. Wood is ground to fibers for hardboard, medium-density fiberboard, and paper products. It is sliced and flaked for

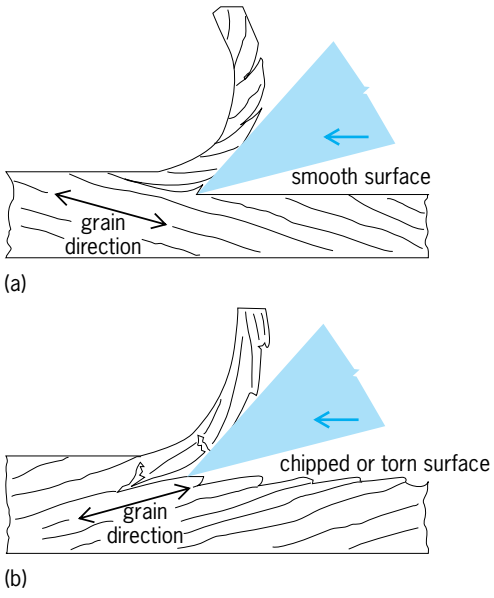


Fig. 3. Cutting direction and wood grain. (*a*) Cutting with the grain leaves a very smooth surface. (*b*) Cutting against the grain leaves a chipped surface. (*After R. B. Hoadley, Understanding Wood, Taunton Press, 1981*)

particleboard products, including wafer boards and oriented strand boards. Whether made from waste products (sawdust, planer shavings, slabs, edgings) or roundwood, the individual particles generally exhibit the anisotropy and hygroscopicity of larger pieces of wood. The negative effects of these properties are minimized to the degree that the three wood directions (longitudinal, tangential, and radial) are distributed more or less randomly. *See* WOOD PRODUCTS.

Charles J. Gatchell

Bibliography. J. G. Haygreen and J. L. Bowyers, *Forest Products and Wood Science: An Introduction*, 3d ed., 1996; R. B. Hoadley, *Understanding Wood: A Craftsman's Guide to Wood Technology*, 2d ed., 2000; U.S. Department of Agriculture, *Dry Kiln Operator's Manual*, Agr. Handb. 188 (rev.), 1991; U.S. Department of Agriculture, *Wood Handbook: Wood as an Engineering Material*, Agr. Handb. 72 (rev.), 1974.

Wood products

Products, such as veneer, plywood, laminate of products, particleboard, waferboard, pulp and paper, hardboard, and fiberboard, made from the stems and branches of coniferous (softwood) and deciduous (hardwood) tree species. The living portion of the tree is the region closest to the bark and is commonly referred to as sapwood; the dead portion of the tree is called heartwood. In many species, especially hardwoods, the heartwood changes color because of chemical changes in it. The heartwood of walnut, for example, is dark brown and the sapwood, almost white.

Wood is one of the strongest natural materials for its weight. A microscopic view reveals thousands of hollow-tubed fibers held together with a chemical called lignin. These hollow-tubed fibers give wood its tremendous strength for its light weight. These fibers, after the lignin bonding material is removed, make paper. In addition to lignin, wood is composed of other chemicals, including cellulose and hemicellulose. *See* CELLULOSE; LIGNIN; WOOD ANATOMY.

Solid wood products. This category includes lumber products, veneer, and structural plywood.

Lumber. Most small log sawmills try to maximize value and yield by automating as much of the manufacturing process as possible. The basic process of cutting lumber involves producing as many rectangular pieces of lumber as possible from a round tapered log. There are only a few sawing solutions, out of millions possible, that will yield the most lumber from any given log or larger timber.

The first stage of the process involves bucking (cross-cutting) the logs into sawmilllength logs (8–24 ft or 2.4–7 m). If the tree-length logs (which can reach a maximum length of 65–70 ft or 20–21 m) are longer than what can be transported over the roadways, the bucking is done in the forest. Otherwise, the bucking is done at the sawmill. In many cases, the logs are electronically scanned for geometric shape and, based on lumber market values, the logs are bucked to optimum length by computer. After bucking, the logs are ready for primary breakdown, or cutting on the first, or headrig, sawing machine. The goal of the headrig is to cut the log into as few pieces as possible and pass them to a machine center downstream in the flow. The log is first scanned for geometric shape, and a computer determines how best to saw it. The log is automatically positioned, and the headrig saws are automatically set. The log then is processed into a combination of side boards and a larger center cant which is cut further downstream. In some areas, a headrig may consist of chipping devices that convert the rounded outside strips of log directly into chips for paper instead of cutting the strips with a saw. After processing through the headrig, the side boards are processed through a board edger. This computerized machine removes the rounded edges (wane) from the sides of the board. The side boards are scanned by lasers and precisely edged by circular saws.

The center cants from the logs are processed through a computerized sawing machine called a rotary gang saw; the machine uses circular saws, and there are up to eight saws spaced evenly apart on a central rotating shaft (arbor). These machines are among the most accurate in the sawmill. Because the center cant makes up a very large proportion of the volume of the log, accuracy is very important.

Following edging and cant processing, the lumber is trimmed to accurate length. A computerized trimmer scans the boards for shape and surface defects and then trims them minimally. Some boards may be reprocessed through a resaw prior to trimming. The resaw uses a band saw to remove any irregular shapes from lumber. Following the trimming operation, the lumber leaves the sawmill, where it is processed through an automated sorting and stacking machine. Each piece of lumber is electronically measured and transported to a bin where boards of similar dimension are stored. When enough lumber of the same size is accumulated, the lumber is stacked for air drying, kiln drying, or planing. The planing process puts a smooth surface on the faces and edges of the board. Following planing, the lumber is graded for structural or appearance characteristics, then packaged for shipment.

Veneer and structural plywood. Structural plywood is constructed from individual sheets of veneer, often with the grain of the veneer in perpendicular directions in alternating plies. The most common construction is three-, four-, and five-ply panels. The alternating plies give superior strength and dimensional stability.

Plywood is used in many applications, including roof sheathing, siding, and floor underlayment. The veneer is rotary peeled from short logs (called blocks) which are slightly over 8 ft (2.4 m) in length. The process begins by bucking the logs into plywood block lengths (8–9 ft or 2–3 m). This process, as with sawmilling, may be computerized. Following bucking, the blocks are usually steam conditioned prior to peeling so that the wood does not split and the knots are softened.

The plywood blocks are electronically scanned and precisely positioned in the veneer lathe. The lathe peels the veneer from the log with a veneer knife. The veneer leaves the lathe and proceeds to a computer automated veneer clipper that clips out the defects (such as knots and splits). The types of veneer produced by the clipper are full, 4×8 ft (1.2×2.4 m) sheets; half, 2×8 ft (0.6×2.4 m) sheets; and strip, less than 2 ft (0.6 m) in width. The full and half sheets are automatically stacked, and the strip is hand stacked and stored to await drying.

The plywood veneer is segregated into heartwood and sapwood prior to drying. The sapwood has much more moisture in it, and requires longer drying times than heartwood. Drying times range 5–15 min. After drying, the veneer is again stacked to await panel lay-up and gluing. The individual pieces of veneer are then either manually or semiautomatically processed through a lay-up line. The lay-up line places the veneer in its proper orientation and applies adhesive to each piece. Often, the cross plies are placed on the line manually.

After the lay-up process, the panels are loaded into a hot press where the temperature is raised to polymerize the adhesive. The panels are removed from the press and proceed through a patching process where defects are routed out of the panels and a synthetic liquid patch material is applied. This patch material also polymerizes because of the heat of the panel or external heaters. Finally, the panels are trimmed up square, and sometimes sanded and grooved for siding. The panels are then graded and packaged for shipment to the customer.

Composite products. Such products include laminated products made from lumber, particleboard, waferboard, and oriented strand board.

Laminated products. Laminated products are composite products, made from lumber, parallel laminated veneer, and sometimes plywood, particleboard, or other fiber product. The most common types are the laminated beam products composed of individual pieces of lumber glued together with a phenol resorcinol–type adhesive. Laminated beams are constructed by placing high-quality straight-grained pieces of lumber on the top and bottom, where tension and compression stresses are the greatest, and lower-quality lumber in the center section, where these stresses are lower.

Another form of composite beam is constructed from individual members made from parallel laminated veneer lumber. This type of lumber has the advantage that it can be made into any length, thus creating beams to span large sections. These beams can be made entirely of parallel laminated veneer lumber, or with the top and bottom flange made from parallel laminated veneer lumber, and the center web made from plywood or flakeboard. Some of these products use solid lumber for the flange material and resemble an I beam. All structural laminated products have the advantage over lumber in that much of the natural variation due to defects is removed, and wood structural members can be made much larger than the typical 2×12 in. (51×305 mm) lumber product.

Particleboard. Nonstructural particleboard is another type of composite product that is usually made from sawdust or planer shavings. It is sometimes made from flaked roundwood. This type of particleboard is one of the most widely used forms of wood product. Often hidden from view, it is used as a substrate under hardwood veneer or plastic laminates. It is commonly used in furniture, cabinets, shelving, and paneling. Particleboard is made by drying, screening, and sorting the sawdust and planer shavings into different size classifications. This particleboard (furnish) is then placed in an adhesive blender where a urea-formaldehyde adhesive is applied. Following blending, the furnish is transported to a forming machine where the particleboard mat of fibers is created. Commonly, the forming machine creates a loose mat consisting of fine particles on the outside and coarser particles on the inside. After the mat is formed, it is transported to a multiopening hot press where it is compressed into the final panel thickness and the adhesive is cured. Following pressing, the large panels are cooled, trimmed, and cut into smaller finished sizes. They are next sanded, and graded prior to shipment to furniture and other manufacturers.

Waferboard and oriented strandboard. Waferboard and oriented strandboard are structural panels made from flakes or strands, and are usually created from very small trees. Unlike nonstructural particleboard, waferboard is designed for use in applications similar to those of plywood. Waferboard and oriented strandboard can have flakes or strands oriented in the same direction, thus giving the board greater strength in the long axis. The most common type of waferboard is made from randomly oriented flakes. The process of making waferboard is similar to that of particleboard except that the wood material used is in the form of flakes instead of particles. The adhesive is a phenol formaldehyde which is usually applied in a powdered form.

Fiber products. The most common fiber products result from pulping processes that involve the chemical modification of woodchips, sawdust, and planer shavings. Such products include pulp and paper, hardboard, and fiberboard.

Pulp and paper. Paper making begins with the pulping process. Pulp is made from wood chips created in the lumber manufacturing process, small roundwood which is chipped, and recycled paper. The fibers in the chips must be separated from each other, by mechanically grinding the fibers or chemically dissolving the lignin from them. The most common chemical processes are sulfite and sulfate (kraft). Following the pulping process, the fibers are washed to remove pulping chemicals or impurities. In some processes (for example, writing papers), the fibers are bleached.

During the beating and refining process, the fibers are flattened to create more fiber bonding sites. This process is necessary to give paper added strength. The fibers are suspended in a water slurry and processed through a Fourdrinier machine. During the wet stage, the water is removed from the pulp, and

the resulting mat of fibers is transported on a screen to the wet pressing and drying section, where the mat is passed through a series of hot rollers. Dry, finished paper emerges from the end of this section, and it is placed in rolls for further manufacture into paper products. *See* PAPER.

Hardboard. Hardboard is a medium- to high-density wood fiber product made in sheets from $^1/_{16}$ to $^1/_2$ in. (1.6 to 12.7 mm). Hardboard is used in furniture, cabinets, garage door panels, vinyl overlaid wall panels, and pegboard. It is made by either a wet or dry process. Fibers are produced from chips through a thermomechanical process which retains the lignin that is used to bond the fibers together. During the wet process, the pulp is mixed with water and placed on a screen where the water drains away and the pulp mat is retained. The mat is placed in a prepress, where the excess water is removed. The fiber mat, still on the screen, is then placed in a hot press where the panel is formed. During the dry process, air is used instead of water to form the mat prior to pressing. After the hardboard is pressed, it is tempered by using an oil soak process or high-temperature exposure.

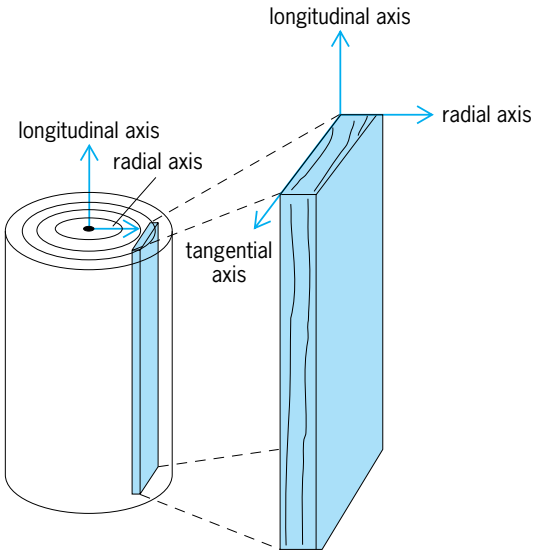
Medium-density fiberboard. Medium-density fiberboard is used in many of the same applications where particleboard is used. It can be used in siding and is especially well suited to cabinet and door panels where edges are exposed. Unlike particleboard which has a rough edge, medium-density fiberboard has a very fine edge that can be molded very well. Medium-density fiberboard is produced in much the same way that dry-processed hardboard is produced in its early stages. The chips are thermomechanically pulped or refined prior to forming into a dry mat. Following refining, medium-density fiberboard is produced in a fashion similar to that of particleboard. The dry pulp is sprayed with adhesive (usually urea-formaldehyde or phenol-formaldehyde) and formed into a dry mat prior to pressing in a multiopening hot press. After the panels have been formed, they are cooled and cut to smaller final product sizes prior to shipment. *See* WOOD PROCESSING.

Terence D. Brown

Bibliography. R. F. Baldwin, *Plywood Manufacturing Practices*, 1984; C. J. Biermann, *Essentials of Papermaking*, 1993; T. D. Brown, *Quality Control in Lumber Manufacturing*, 1982; J. G. Haygreen and J. L. Boywer, *Forest Products and Wood Science*, 3d ed., 1996; T. M. Maloney, *Modern Particleboard and Dry Process Fiberboard*, 1972, updated 1993; E. M. Williston, *Lumber Manufacturing*, 1988.

Wood properties

Physical and mechanical characteristics of wood which are controlled by specific anatomy, moisture content, and to a lesser extent, mineral and extractive content. The properties are also influenced by wood's directional nature, which results in markedly different properties in the longitudinal, tangential, and radial directions or axes (see **illus.**). Wood properties within a species vary greatly from tree to tree and within a single axis. The physical properties (other than appearance) are moisture content, shrinkage, density, permeability, and thermal and electrical properties.



Coordinate system for a log or section of a tree (left) and a piece of lumber from the tree (right).

Moisture content. Moisture content is a major factor in the processing of wood because it influences all physical and mechanical properties, and durability and performance during use. Normal in-use moisture content of processed wood that has been dried ranges 8–13%. Water is held in wood in three ways: as water chemically bonded to hydroxyl groups in the wood substance of the cell wall, or as either free water or water vapor in the cell cavities. When the free water is removed from the cell cavities and the cell wall is saturated to its maximum, the wood is said to be at its fiber saturation point, which ranges 25–35% moisture content for most species. Wood can chemically adsorb or desorb moisture as water vapor below the fiber saturation point. Because of the capillary structure of the cavities, wood can also mechanically absorb liquid water.

Moisture content for wood is expressed on either a fractional or percentage basis. Moisture content is defined as the ratio of the mass of water contained in the wood to the mass of the same sample of dry wood. For hardwoods in the green state, the moisture content of sapwood and heartwood is approximately equal, ranging 45–150% according to species. The moisture content of softwoods in the green condition differs for heartwood and sapwood, with heartwood ranging 30–120% and sapwood 100–250% according to species. Thus, the average moisture content of undried softwoods is affected by the percentages of heartwood and sapwood in the sample.

Shrinkage. Shrinkage occurs when wood loses moisture below the fiber saturation point. Above that point, wood is dimensionally stable. The amount of the shrinkage depends on its direction relative to grain orientation and the amount of moisture lost

below the fiber saturation point. Volumetric shrinkage is about 10–19% depending on species. Wood shrinks significantly more in the radial and tangential directions than in the longitudinal direction.

Density. The density of wood is determined by the amount of cell wall substance and the volume of voids caused by the cell cavities (lumens) of the fibers. Density can vary widely across a growth or annual ring. For species with distinct growth rings, the earlywood (light bands on the cross section of wood) consists of fast-growing, relatively thin-walled cells or fibers (that is, less dense wood), and the latewood is composed of thick-walled, slow-growing cells or fibers (that is, more dense wood). The percentage of earlywood and latewood in each growth ring determines the overall density of a wood sample.

Permeability. Permeability is a measure of the flow characteristics of a liquid or gas through wood as a result of the total pressure gradient. Permeability is influenced by the anatomy of the wood cells, the direction of flow (radial, tangential, and longitudinal), and the properties of the fluid being measured. The longitudinal flow is greater than the flow in either the radial or tangential directions. The longitudinal flow is relatively unrestricted through the cavities of the cells, while the tangential and radial flows are affected by the quantity, size, and condition of pits in the cell wall. Permeability is also affected by the species, by whether the wood is sapwood or heartwood, and by the chemical and physical properties of the fluid.

Thermal properties. The primary thermal properties of wood are conductivity, specific heat, and coefficient of thermal expansion. When wood is ovendry it is a poor conductor of heat, whereas when wood has a high moisture content it is an excellent conductor. The conductivity of wood is determined by density, moisture content, and direction of conduction. Thermal conductivity in the transverse directions (radial and tangential) is approximately equal. Conductivity in the longitudinal direction is greater than in the transverse directions. The greater conductivity in the longitudinal direction is due to cell walls which are less interrupted than in the transverse directions.

For most processing operations, the dominant heating direction is transverse. Thermal conductivity is important to wood processing because heating—whether for drying, curing, pressing, or conditioning—is an integral step.

Specific heat of wood is dependent on moisture content and, to less extent, on temperature. *See* SPECIFIC HEAT.

Electrical properties. Dry wood is an excellent insulator. By measuring wood's electrical resistance, electrical moisture meters accurately determine the moisture content of wood in the 5–25% range. Two other electrical properties of interest are the dielectric constant and the dielectric power factor for alternating current. These dielectric properties are dependent on density, moisture content, frequency of current, grain orientation, and temperature. The power factor is a measure of the stored energy that is converted to heat. This power factor is also affected by frequency, moisture content, and temperature. The dielectric character of wood is used as a working principle in moisture meters to measure the moisture content of wood over its entire moisture range.

Mechanical properties. The mechanical properties of wood include elastic, strength, and vibration characteristics. These properties are dependent upon species, grain orientation, moisture content, loading rate, and size and location of natural characteristics such as knots.

Because wood is an orthotropic material, it has unique and independent mechanical properties in each of three mutually perpendicular axes—longitudinal, radial, and tangential. This orthotropic nature of wood is interrupted by naturally occurring characteristics such as knots that, depending on size and location, can decrease the stiffness and strength of the wood.

Elastic properties. Wood is both an elastic and plastic material. Elasticity manifests itself during loading and at moisture contents and temperatures that occur in most service uses of wood. Wood is primarily elastic to its proportional limit during loading at room temperature, at which point the wood becomes plastic and starts an irrecoverable flow. The elastic stiffness or modulus of elasticity of wood is dependent on grain orientation, moisture content, species, temperature, and rate of loading.

The stiffness of wood in the longitudinal (fiber) direction is utilized in the manufacture of composite products such as oriented strand board, in which the grain or fiber direction is controlled. The stiffness of wood may be predicted by using the ovendry density of the wood and a correction factor for moisture content. In industrial applications, the measured stiffness of lumber can be used to predict its strength rating nondestructively. *See* ELASTICITY.

Strength. The strength of wood, like its elastic properties, is dependent upon rate of loading, species, moisture content, orientation, temperature, size and location of natural characteristics such as knots, and specimen size. The strength of individual wood fibers in the longitudinal direction can be significantly greater than that of larger samples with their complex anatomy and many defects. As with stiffness, the excellent strength characteristics of wood in the direction of the fiber can be maximized during the manufacture of wood composites by controlling fiber alignment.

Vibration. Damping and sound velocity are two primary vibration phenomena of interest in structural applications. Damping occurs when internal friction dissipates mechanical energy as heat. The damping characteristic of wood results in quick dissipation of all energy for most frequencies of vibration. Because of its high damping capacity, wood is an excellent material in floors and other structural components, and it is excellent for resisting the forces resulting from earthquakes.

The velocity of a sound wave through wood can be used to estimate mechanical stiffness and strength:

the higher the velocity, the higher the stiffness and strength. Like other properties of wood, the velocity of sound along the three principal axes differs. The three directional velocities are proportional to the wood's stiffness values in these directions. Sound velocity in the longitudinal direction is two to four times greater than in the transverse directions. *See* WOOD ANATOMY.

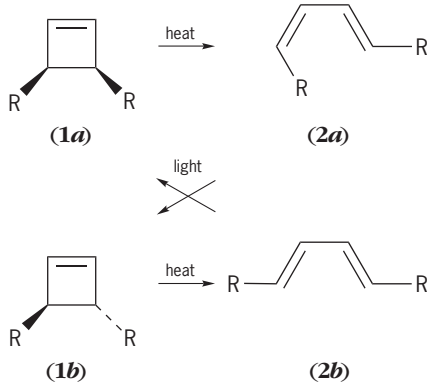
Jame B. Wilson

Bibliography. J. Bodig and B. A. Jayne, *Mechanics of Wood and Wood Composites*, 1982; Forest Products Laboratory Staff, *Wood Handbook: Wood as an Engineering Material*, 1987, revised 2000; J. F. Siau, *Transport Processes in Wood*, 1984; C. Skaar, *Wood-Water Relations*, 1988.

Woodward-Hoffmann rule

A concept which can predict or explain the stereochemistry of certain types of reactions in organic chemistry. It is also described as the conservation of orbital symmetry, and is named for its developers, R. B. Woodward and Roald Hoffmann. The rule applies to a limited group of reactions, called pericyclic, which are characterized by being more or less concerted (that is, one-step, without a distinct intermediate between reactants and products) and having a cyclic arrangement of the reacting atoms of the molecule in the transition state. Most pericyclic reactions fall into one of three major classes, examples of which will illustrate the use of the rule. *See* PERICYCLIC REACTION.

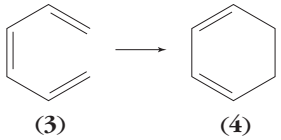
Electrocyclic reactions. These reactions are defined as the interconversion of a linear π system, containing n π electrons, and a cyclic molecule containing $(\eta - 2)$ π electrons which is formed by joining the ends of the linear molecule. This is exemplified by the thermal ring opening of a cyclobutene to a butadiene, reaction (1).



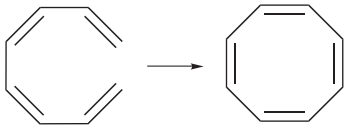
(**1*a***) (**2*a***) (1)

(**1*b***) (**2*b***)

The reaction is stereospecific in that (**1*a***; R substituents cis) gives only (**2*a***) and none of (**2*b***). Conversely, cyclobutene (**1*b***) gives (**2*b***) but not (**2*a***). This mode of reaction is termed conrotatory, since both R groups rotate in the same direction (clockwise or counterclockwise when viewed edge-on) in going from reactant to product. On irradiation with ultraviolet light the butadienes recyclize to cyclobutenes. This reaction is disrotatory since (**2*a***) → (**1*b***) and (**2*b***) → (**1*a***). Similar stereospecificity is observed in the cyclization of substituted hexatrienes (**3**) and octatetraenes (**5**) to cyclohexadienes (**4**) and cyclooctatrienes (**6**), respectively [reactions (2) and (3)].



(**3**) (**4**) (2)



(**5**) (**6**) (3)

To see how these results are explained requires a knowledge of the nature, specifically the symmetry, of the molecular orbitals most involved in the reaction, such as the relevant occupied orbitals of butadiene and cyclobutene (**Figs. 1** and **2**). The signs (+ or −) in the orbitals indicate the phase of the wave function in that region of space. *See* MOLECULAR ORBITAL THEORY.

The orbitals are labeled S or A according to their symmetry with respect to rotation by 180° about an axis in the plane of the molecules (Fig. 1). This

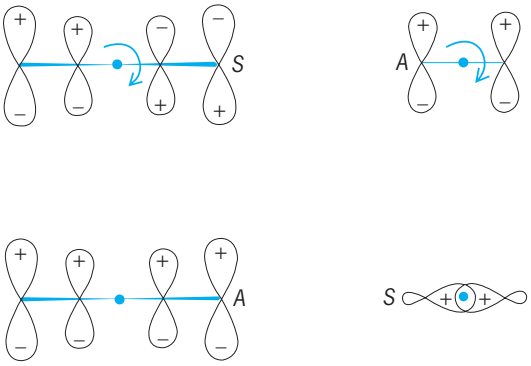


Fig. 1. Conrotatory cyclization, butadiene: axial symmetry.

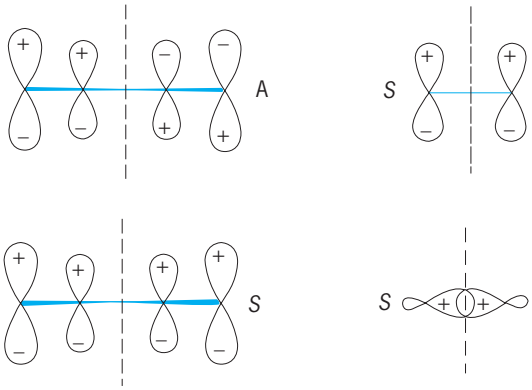


Fig. 2. Disrotatory cyclization, cyclobutene: planar symmetry.

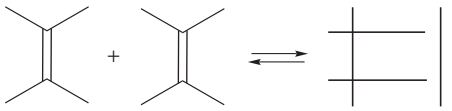
Electrocyclic reactions of polyenes

n	Thermal reaction	Photochemical reaction
$4q$ (q = 4, 8, . . .)	Conrotatory	Disrotatory
$4q + 2$ (q = 2, 6, . . .)	Disrotatory	Conrotatory

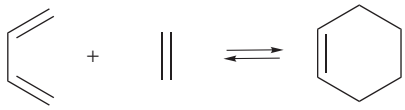
element of symmetry is maintained throughout the conrotatory cyclization. The same orbitals and their symmetry with respect to reflection through the plane that bisects the molecules holds for cyclobutene (Fig. 2). The molecule retains this element of symmetry in its disrotatory cyclization. The fact that in the conrotatory mode of cyclization both butadiene and cyclobutene have the same number of symmetric occupied orbitals (one each) shows that the reaction can take place; that is, it is symmetry-allowed. In the disrotatory mode there are two occupied symmetric orbitals in cyclobutene, but only one in butadiene. Thus in the latter path, orbital symmetry is not conserved, and the reaction is said to be symmetry-forbidden.

The general rules for electrocyclic reactions of polyenes with n π electrons are summarized in the **table**, where q is an integer.

Cycloaddition reactions. The simplest example of this type of pericyclic reaction is the combination of two unsaturated molecules, ends to ends, to form a cyclic molecule with four fewer π electrons, for example, reactions (4) and (5). When these



(4)

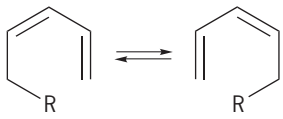


(5)

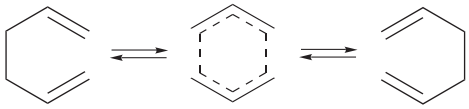
reactions are concerted, the Woodward-Hoffmann rule correctly predicts the experimentally observed stereochemistry. A generalization which can be made is that cycloadditions of two or more unsaturated molecules are most facile when there are a total of ($4q + 2$) π electrons ($q > 0$). Cycloadditions (or the reverse reactions) involving $4q$ electrons ($q > 0$) are symmetry-allowed, but the stereochemistry of the required transition state is usually sufficiently strained that the reaction proceeds in two steps.

Sigmatropic reactions. This type of reaction is best defined by reference to the examples shown in reactions (6)–(8). Note that in each reaction there is no

(6)



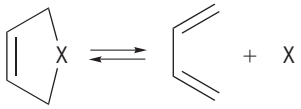
(7)



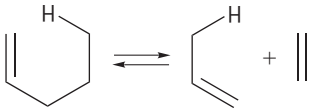
(8)

overall change in the number of σ or π bonds. The name sigmatropic is derived from the fact that there is a change in the location of the σ bond connecting the upper and lower fragments of the molecule. The individual reactions are classified in terms of the number of atoms from each fragment in the cyclic transition state. Thus the above examples are sigmatropic reactions of order [1,3], [1,5], and [3,3], respectively. All of the above and many other such reactions are known, and in each case the Woodward-Hoffmann rule predicts the experimentally observed stereochemistry.

There are several other miscellaneous pericyclic reactions to which orbital symmetry theory applies. Among these are the cheletropic reactions, for example, (9) and the Ene reaction (10). Although the



(9)



(10)

examples have shown simple hydrocarbons as reactants and products, the rules apply also to molecules containing atoms other than carbon and hydrogen. A methylene group may be replaced by an oxygen or nitrogen atom, for example, and the conclusions from orbital symmetry are unaltered.

It should be recognized that the primary explanation for the occurrence of any chemical reaction is found in either the strengths of the reactant's or the products' bonds or both. If the reaction is pericyclic, the conservation of orbital symmetry can affect the rate of the reaction. Thus if two reactions or two mechanisms for the same reaction are possible, the one which is symmetry-allowed will be much faster than the one which is symmetry-forbidden. If, for reasons unrelated to orbital symmetry, no symmetry-allowed reaction is possible for an apparently pericyclic reaction, the reaction may be observed to proceed via a forbidden path. Usually, however, it is found on careful study that the reaction avoids orbital symmetry control by proceeding in two steps via a noncyclic intermediate; that is, it is not pericyclic. *See* ORGANIC REACTION MECHANISM; STEREOCHEMISTRY.

David L. Dalrymple

Bibliography. S. N. Ege, *Organic Chemistry: Structure and Reactivity*, 4th ed., 1999; E. A. Halevi, *Orbital Symmetry and Reaction Mechanism: The OCAMS View*, 1992; R. Hoffmann and R. B. Woodward, The conservation of orbital symmetry, *Account. Chem. Res.*, 1:17, 1968; P. Laszlo, *Organic Reactions: Simplicity and Logic*, 1996; R. B. Woodward and R. Hoffmann, *The Conservation of Orbital Symmetry*, 1970.

Woodworking

The shaping and assembling of wood and wood products into finished articles such as mold patterns, furniture, window sashes and frames, and boats. The pronounced grain of wood requires modifications in the working techniques when cutting with the grain and when cutting across it. Five principal woodworking operations are sawing, planing, steam bending, gluing, and finishing. To shape round pieces, wood is worked on a lathe. *See* TURNING (WOODWORKING).

Sawing. Wood is sawed by cutting or splitting its fibers by the continuous action of a series of uniformly spaced teeth alternately staggered to move in closely parallel work planes. Action of the cutting teeth produces a path or kerf of uniform width through the workpiece from which the fibers have been severed and removed. Sawing across the grain or cell structure of the wood is called crosscutting. Cutting parallel with the grain of the piece is referred to as ripping. Saw teeth are bent alternately to the left and right to provide clearance for the blade. Some blades include straight raker teeth for cleaning fibers from the cut.

Woodcutting saws may be classed as either handsaws or power-operated saws. Either group consists of numerous types and designs.

Handsaws. Each type of handsaw is designed to accomplish one specific type of sawing operation most effectively.

Crosscut handsaws are made with about 8 teeth per inch of length. Ripsaws for cutting with the grain usually have about $5^1/_2$ teeth per inch (2 per centimeter). Finetooth saws for finishing or cabinetwork may have as many as 10–16 teeth per inch (about 4–6 per centimeter) of length.

A backsaw is a fine-tooth saw with its upper edge stiffened to ensure straight cuts. Keyhole saws with their narrow, tapered blades are used for cutout work where sharp turns are required. A compass saw has a handle with several attachable blades of varying widths, making it suitable for a variety of work. Coping saws have narrow blades usually about $^1/_8$ in. (0.3 cm) wide. The blade is held taut in a frame which is equipped with a handle. The narrow blade and high-backed frame make the saw suitable for shaping or cutout work.

Power saws. Power-operated woodworking saws are usually combined with auxiliary equipment that enables them to perform various sawing operations.

Bench or circular saws are the common woodworking type of power saw (**Fig. 1**). Depending on construction, either a power arbor or worktable may be raised, lowered, or tilted. The saw may be used for crosscutting, ripping, or resawing, and for beveled as well as tapered cuts. Molding cutters along with those designed for making rabbet, tendon, and dado joints are also used. Accessories for sanding, buffing, and polishing are available for most models.

Band saws are basically a flexible band of steel running over two vertical pulleys (**Fig. 2**). The band or blade has teeth on one side and is operated under tension. The wide distance or throat between the cutting portion of the blade and the rear blade guide and support arm adapt the band saw for cutout work or sawing on large flat pieces.

Scroll saws are used for work similar to that performed by the band saw. The continuous-band type of blade is replaced by a short, vertically reciprocating blade (**Fig. 3**).

Radial saws have their circular blade mounted above the worktable. The blade and motor are suspended from an overarm that allows travel across the workpiece: a pivot permits cuts to be taken at any angle (**Fig. 4**). Usually the saw may be raised and lowered as well as tilted at an angle. Crosscutting, ripping, mitering, and beveling may be performed; accessories and attachments permit other circular cutting tool operations such as dadoing, molding, drilling, and sanding.

Portable handsaws consist of a circular blade and electric motor plus the necessary frame, handles, baseplate, and guards. An electric cord of reasonable length permits the saw to be manually moved and positioned for the desired cut. Some models may be fastened to a special frame or table.

Portable saber saws are compact units consisting of an electric motor, a straight saw blade driven by a

Fig. 1. Bench circular saw with tilting arbor is used for parting or slotting, and can make cuts as long as working space permits. (*Delta International Machinery Corp., Pittsburgh*)

Fig. 2. Narrow bandsaw, a flexible band of steel, can make curved as well as straight cuts even in thick pieces. (*Delta International Machinery Corp., Pittsburgh*)

Fig. 3. Scroll saw, a short, vertically reciprocating blade, cuts sharp turns in thin pieces. (*Delta International Machinery Corp., Pittsburgh*)

Fig. 4. Radial saw allows wood to be clamped in position while it is being cut. (*Delta International Machinery Corp., Pittsburgh*)

reciprocating mechanism, handle, baseplate, electric cord, and other necessary parts. The lightness of the saw and its narrow blade fastened at only one end make it adaptable to many types of cutting including cutout work and shaping.

Planing. Flat or uniformly contoured surfaces of wood are roughed down, smoothed, or made level by the shaving and cutting action of a wide-edged blade or blades. Planing may be accomplished either manually or by power-operated tools.

Hand planes. Manually operated planes are classified as either bench or block types. A bench plane is used for shaving with the grain of the wood, whereas a block plane is designed for cutting across the grain. A block plane is usually small; bench planes vary in size and type. The common bench types are the smoothing, jack, fore, and jointer.

The smoothing plane, 5.5–10 in. (14–25 cm) long, is best suited for smoothing small areas. The

Fig. 5. Six-inch-long (15-cm) bed, jointer planer found in woodworking shops is used for surfacing, rabbeting, beveling, tapering, molding, and cutting round tenons. (*Delta International Machinery Corp., Pittsburgh*)

somewhat larger jack plane may be used for roughing down or leveling. Fore and jointer planes are still larger in size, with the latter being approximately 22–24 in. (55–60 cm) long. It is used to plane long surfaces.

Special hand planes are the rabbet plane, used to cut recesses for rabbet joints; the model maker's, used to remove excess wood from a curved surface; the scrub or roughing plane, which has heavy, rounded blades making it suitable for cleaning up rough boards; and the circular plane, with a flexible steel bottom that may be adjusted to fit a curved surface.

Power planes. Power-operated planes vary in size and design according to the application and type of work handled. The planer usually found in the woodworking shop is the jointer, frequently called a jointer planer (**Fig. 5**).

The jointer is designed with its cutting blades or knives fastened in a rotating cutterhead. The lengths of the knives classify the machine as to the width of board that it can surface.

Two tables, one in front of the cutterhead and one behind it, support the workpiece as it is pushed through the path of the knives. The front table is set lower than the highest point on the arc of the rotating knives by the amount to be planed from the board. The rear table must be aligned exactly with the high point of the knives so that the workpiece will not pass through the cutting path at an angle.

A tilting fence or guide is provided along one side of the tables. The opposite side of the front table has a ledge which is used to support the workpiece for rabbeting cuts.

Steam bending. Wooden members are bent or formed to a desired shape by pressure after they have been softened or plasticized by heat and moisture. If thick pieces of wood are to be bent to a permanent shape without breaking, some form of softening or plasticizing such as steaming is necessary.

When a piece of wood is bent, its outer or convex side is actually stretched in tension while its concave side is simultaneously compressed. Actually, plasticized wood can be stretched but little. It can, however, be compressed a considerable amount. When a piece of plasticized wood is successfully bent, the deformation is chiefly compression distributed almost uniformly over the curved portion. Curvature results from many minute folds, wrinkles, and slippages in the compressed area.

Steaming. Although soaking wood in water softens it somewhat, a combination of heat and moisture can produce a degree of plasticity approximately 10 times that of dry wood at normal temperatures. Wood need not be steamed to its maximum plasticity in order to be bent. Wood steamed at atmospheric pressure bends in most cases, as does wood steamed at higher pressures. Higher pressures also tend to overplasticize the wood: this results in an increased number of bending failures. Treatment of wood with boiling water has approximately the same effect as saturating it with steam at atmospheric pressure. The boiling water treatment is usually employed only when a portion of a wooden piece requires softening.

Dry wood that has a moisture content of 12% or less must have its moisture increased to approximately 15% to make it suitable for moderate bends. If more severe bends are required, additional moisture must be added to the surface areas. Wood that already has a moisture content of 20–25% needs no further saturation, even for severe bending.

The time required for steaming is directly related to the amount of moisture already present in the wood. In general, dry stock is steamed 1 h per inch (0.4 h/cm) of thickness and green stock, 0.5 h per inch (0.2 h/cm) of thickness. Steaming is usually done in a closed retort suitable for zero-pressure or low-gage-pressure steam. Steam should enter the retort or steam box through water standing in the bottom so that steam in the box will be wet or saturated.

Wood-bending methods may be classed as made without end pressure (free bends) or made with end pressure. On thick pieces, only slight curvatures are feasible by free bending. Bending with end pressure is necessary to obtain the required compression and to prevent tensile failures in moderate or severe bends. The most common method of bending with end pressure is by means of a metal strap fitted with end blocks or clamps. The strap is placed against the convex side of the piece, and bending pressure is applied at the end of the strap by some suitable means (**Fig. 6**). The end blocks must be applying end pressure simultaneously on the wood to prevent tensile stress and also to supply the necessary compression.

Other devices used for bending plasticized wood are variable-position rollers, hot-plate form presses, and special mechanical devices.

When a workpiece is removed from the bending device, it has a tendency to spring back somewhat. This is counteracted by holding the piece in position until it has dried or set. Frequently springback is compensated for by overbending.

Fig. 6. Bent boat rib is prepared for drying with end pressure applied by the strap and the rods placed against the convex side of the piece. Wood stays assist in holding the work in position. (*Forest Products Laboratory*, *Madison*, *Wisconsin*)

The selection of a species of wood to use as a moderately bent member is governed primarily by its suitability and availability. However, if a severe bend is required, the wood must be selected chiefly for its bending qualities. In general, hardwoods are of better bending quality than softwoods. Pieces should be of fairly straight grain and free of knots or other defects. White and red oak, hickory, elm, ash, beech, birch, maple, mahogany, walnut, and sweet gum are species commonly used for bending.

Adhesive bonding and gluing. Wood pieces may be fastened together by the adhesive qualities of a substance that sets or hardens into a permanent bond. Adhesives for wood are of two principal types, synthetic and natural-origin. The term glue was first applied to bonding materials of natural origin, while adhesive has been used to describe those of synthetic composition. The terms are used interchangeably, but adhesive better covers all types of materials in use. *See* ADHESIVE.

Synthetic adhesives include phenolics, ureas, melamines, polyvinyl resin emulsions, hot melts, epoxies, contacts, mastics, and various combinations of specific adhesives. A hardener or setting agent is usually required to convert synthetic adhesives from liquid to solid. These agents may be furnished separately for addition to the resin before use, or they may be already present in the resin as supplied. Since each type of synthetic adhesive has its own characteristics or requirements for use and application, such as temperature and pressure, and for service conditions, they should be studied and considered on an individual basis before use. Some types require special equipment such as heating facilities and presses, which tend to limit them to commercial use. Some thermosetting resins require temperatures of over 300°F (150°C) and pressures as high as 250 lb/in.^{2} (1.7 megapascals) depending on the wood, which tend to limit them to commercial use. *See* POLYVINYL RESINS; UREA-FORMALDEHYDE RESINS.

While adhesives of natural origin, such as animal casein, soybean, starch, and blood glues, are used to bond wood in some plants and shops, they have been replaced largely by synthetics. Animal glue is probably the natural adhesive most widely used, although casein glue is being used a great deal for structural laminating. As with the synthetic adhesives, the conditions for most suitable use and service of the natural adhesives vary, and should also be considered for use on an individual basis.

Prior to any gluing, the mating surfaces should be clean, smooth, and properly fitted or matched. Surfaces should be machined smooth and true, and be essentially free from rough machine marks, chipped or loosened grain, and surface irregularities.

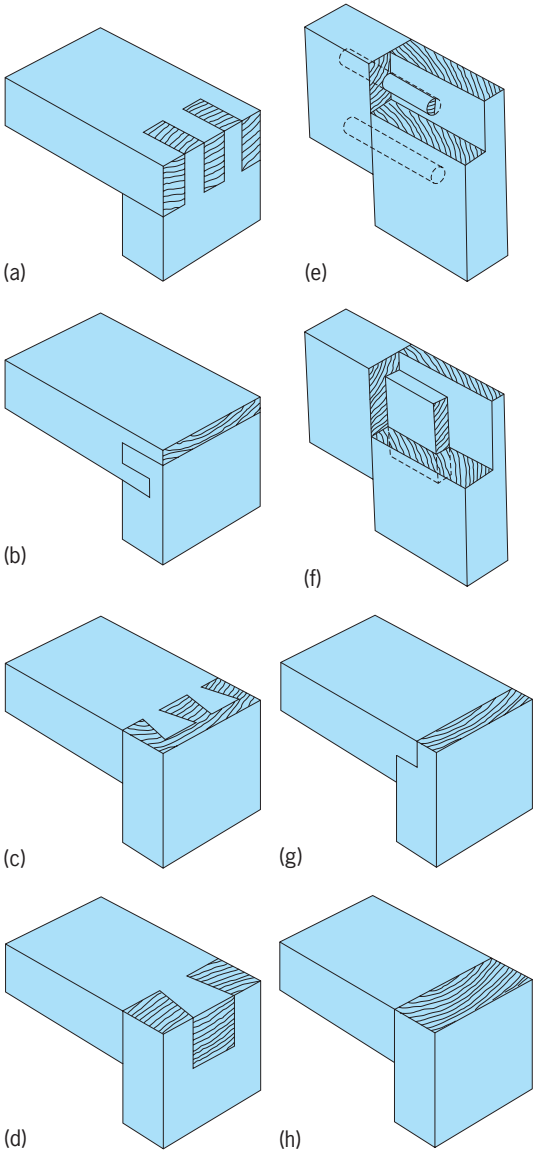


Fig. 7. Various types of corner joints: (*a*) slip or lock corner; (*b*) dado tongue and rabbet; (*c*) blind dovetail; (*d*) dovetail; (*e*) dowel; (*f*) mortise and tenon; (*g*) shouldered corner; (*h*) butt end to side grain. (*U.S. Department of Agriculture*, *Forest Products Laboratory*, *Madison*, *Wisconsin*)

Moisture content of the wood before gluing is important because it affects the quality of the bond and the performance of the glued product in service. Satisfactory adhesion to wood is obtained with most adhesives when the wood moisture content is about 6–17%, and with some glues up to 25%. Glued joints will remain most nearly free from stresses if the moisture content of the parts (when the glue sets) equals the average moisture content that the product will experience in service.

If the end grain surfaces of two pieces of wood are glued together, a butt joint is formed. However, the gluing of square-end butt joints usually does not result in a sufficiently strong and permanent joint to meet the requirements of ordinary or commercial service. Mitered joints (usually cut at a 45° angle with the grain) must also be treated essentially as butt joints for gluing purposes. Seven types of corner joints are shown in **Fig. 7**.

To obtain acceptable strength in pieces spliced together endwise, it is necessary to make a scarf, finger, or other sloped joint (**Fig. 8**). During the gluing operation, end slippage should be prevented to keep parts in proper alignment, and under sufficient and uniform pressure to get maximum joint strength. Even plain scarf joints with a low slope are not as strong as clear wood (of the same quality) in tension parallel to the grain.

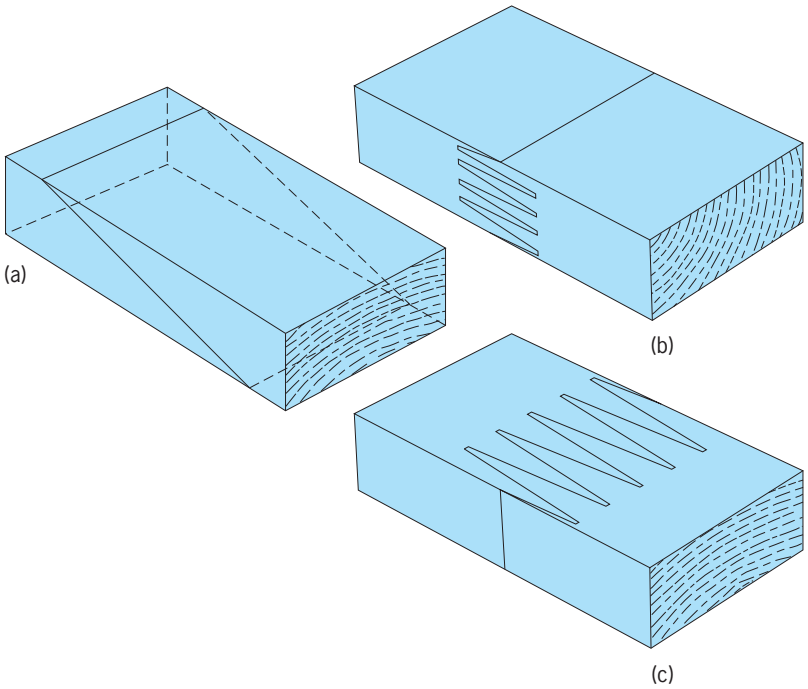


Fig. 8. End joints for splicing lumber: (*a*) scarf joint; (*b*) horizontal finger joint; (*c*) vertical finger joint. (*U. S. Department of Agriculture, Forest Products Laboratory, Madison, Wisconsin*)

Finishing. The finishing operation is the preparation and sealing or covering of a surface with a suitable substance in order to preserve it or to give it a desired appearance. The preparation and conditioning of a surface may include cleaning, sanding, use of steel wool, removing or covering nails and screws, gluing or fastening loose pieces, filling cracks and holes with putty or crack filler, shellacking, and dusting. An inconsequential item with a painted surface does not require the thorough surface preparation that a piece of fine furniture does. The quality of surface conditioning directly affects the end result.

Surface preparation. Wooden surfaces to be painted are cleaned and sanded. Sanding is done with the grain of the wood, using a sandpaper block to keep the surface even, or a power sander (**Fig. 9**). Use of steel wool instead of sandpaper in corners and on rounded surfaces is advisable. Nails and screws should be removed or sunk below the surface, and these holes, as well as any cracks, should be filled with putty or crack filler. New wood is covered with a coating of thinned orange shellac before applying paint. The undercoat tends to prevent the paint from drying unevenly and thus producing flat or glossy spots over knots or other irregularities in the wood.

If the surface is to be varnished, stained, or finished in its natural color, extra care must be taken in sanding, filling in, or smoothing holes and imperfections in the surface. The filler material should be tinted to match the wood. After the filler material has hardened, the entire surface should be sanded with coarse sandpaper. When the rough spots have been removed, sanding should be finished with a No. 1 sandpaper. On fine hardwood surfaces or on furniture from which the varnish has been removed, only fine grades of sandpaper running from No. 0 down to No. 000 should be used. Rubbing must be with the grain, and grit and dust must be removed frequently. If paint remover has been used, the surfaces should be thoroughly cleaned with turpentine and then washed with hot water and soap to remove any wax.

Finish application. Methods used to apply finish coverings to wooden surfaces vary with the substance. In any case, the surface should be wiped clear of dust just prior to applying the material. The

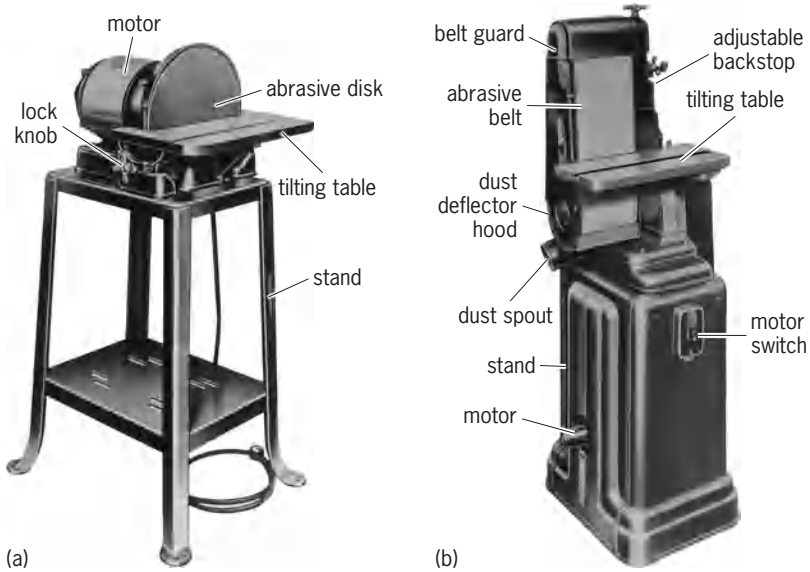


Fig. 9. Types of power sander: (*a*) disk; (*b*) belt. (*Delta International Machinery Corp., Pittsburgh*)

atmosphere and equipment used should be as dust-free as possible.

Paint may be applied by brush or spray as the situation permits. The priming coat and each additional one should be thoroughly dry before applying the next one. On woodwork or painted furniture after each coat prior to the final one, the surface should be sanded lightly to remove any dust or particles stuck to the paint. If a brush is used, the paint should be applied with long even strokes. A thin coat goes on more evenly and tends to prevent the paint from running or forming lumpy spots at the corners and edges.

Stains penetrate the pores of wood but do not fill or close them. Oil stains are the ones most commonly used except on furniture, where water stains are the rule. A few moments after a stain has been applied to a wood surface, any surplus should be wiped off with a lint-free cloth. The time that a stain remains on the wood before being wiped off determines the resulting shade. Although shellac or varnish may be applied over a stain, usually a liquid or paste filler is used to seal the pores in the wood. If varnish is used, two or three coats are generally required.

Varnishes should be applied to a surface quickly and liberally. After most of the varnish in the brush has been applied on an area, it should be spread out evenly. Strokes should be with the grain, and all marks or laps should be brushed out with a nearly dry brush. Varnish used on the first coat should be thinned with about a quart of turpentine to 4 quarts of varnish.

Each coat of varnish should be allowed to dry for several days before applying the next. Sanding should be done with No. 000 sandpaper. Thorough removal of dust is a necessity. Varnishing should be attempted only at temperatures of 65°F (18°C) or above.

Shellac, which is thinner than varnish, dries quickly and must be applied to a surface rapidly. All parts of the surface should be carefully covered with the shellac. Several hours should be allowed for drying between coats. White shellac is used for light woods with a natural finish, while orange shellac is used on all other types of wood. In general, shellac should be used with the same procedure that applies to varnish.

Linseed oil or paste wax is frequently used in finishing furniture. Rubbing may include the use of pumice and alcohol. *See* WOOD PROPERTIES. Alan H. Tuttle

Bibliography. Forest Product Laboratory, *Wood Handbook*, USDA Forest Service, Agr. Handb. 72, rev. August 1974; M. Selbo, *Adhesive Bonding of Wood*, Tech. Bull. 1512, USDA Forest Service, August 1975.

Wool

A textile fiber made of the undercoat of various animals, especially sheep; it may also be obtained from angora, goat, camel, alpaca, llama, and vicuna. Wool provides warmth and physical comfort that cotton and linen fabrics cannot give.

Sheep are generally shorn of their fleeces in the spring, but the time of shearing varies in different parts of the world. Sheep are not washed before shearing. Sometimes they are dipped into an antiseptic bath, but only when prescribed by law. Formerly, sheep were shorn by hand, but today the fleeces are usually removed in one piece by machine clippers, which shear more closely and faster than hand clippers. Wool shorn from young sheep differs in quality from that of older sheep, and wool from live sheep is different from that of dead sheep.

In the United States domestic wool reaches the mill in loosely packed bags; imported wool comes in tightly compressed bales. Each fleece contains different grades, or sorts, of wool, and the raw stock must be carefully graded and segregated according to length, diameter, and quality of fiber. Wool from different parts of the body of the lamb differs greatly. The shoulders and sides generally yield the best quality of wool, because the fibers from those parts are longer, softer, and finer.

Wool technology is that branch of animal science concerned with investigating the structure, growth characteristics, and chemical properties of wool affecting and determining commercial use. The term wool covers the fibers of sheep, angora goats, camels, alpacas, llamas, and vicunas. In this article, however, wool technology refers to sheep fiber only.

Fiber structure. Wool is epidermal in origin, a complex, organized structure growing from a follicle buried in the dermis of the skin (**Fig. 1**). Associated with the follicle are two glands, the sebaceous and the sudiferous. The sebaceous gland secretes wool grease which, when refined, is lanolin. The sudiferous gland secretes sweat, or suint. Collectively, grease and suint in the raw fleece are called yolk, a substance which is widely used in the pharmaceutical and cosmetic industries for lanolin compounds because it can be absorbed by the human skin. Both

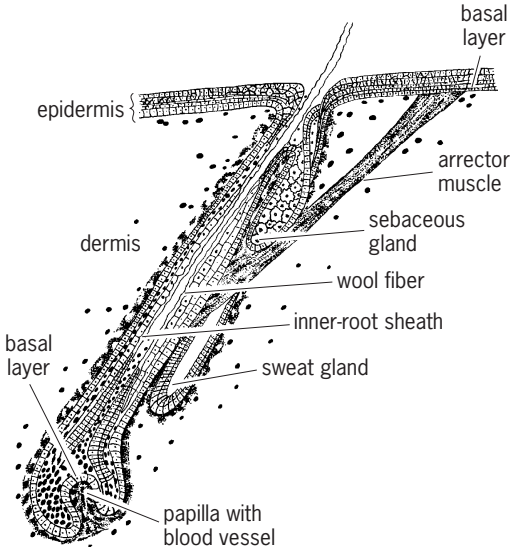


Fig. 1. Longitudinal section of a completely developed nonmedullated wool follicle, including sebaceous and sweat glands.

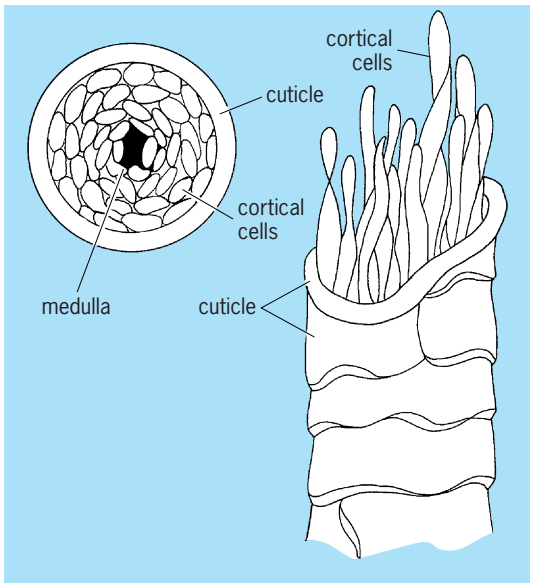


Fig. 2. Cuticle, cortex, and medulla of wool fiber.

glands open into the follicle. Their secretions function as lubricants and protectants for the wool fiber as it grows.

Growth. Blood capillaries in the papilla nourish the growing fiber, which consists of two parts, the root and the shaft. The root, or living part, is beneath the skin surface. The shaft, which protrudes from the mouth of the follicle, is dead. Physical and chemical differences between the root and shaft are listed in **Table 1**.

TABLE 1. Physical and chemical differences between root and shaft of wool fiber*

Root	Shaft
Soft and easily crushed	Tough and horny
Cells roundish	Cells elongated
Positive test for nucleic acid	Negative test for nucleic acid
Nuclei stained with hematoxylin	Nuclei unstained with hematoxylin
Cytoplasm granular in appearance	Cells distinctly fibrous
Not birefringent	Birefringent
Positive test for sulfhydryl groups	Negative test for sulfhydryl groups
No Allwoerden reaction with chlorine water	Many large Allwoerden "sacs"

*After J. M. Matthews and H. R. Mauersberger, *Textile Fibers*, 6th ed., Wiley, 1954.

A cross section of the shaft (**Fig. 2**) reveals three layers: the cuticle, the cortex, and the medulla. The last normally is absent in fine wools and infrequent in improved medium wools.

The cuticle, or epidermis, of the fiber is scalelike, and has overlapping serrated cells, the free ends of which point toward the tip of the fiber (**Fig. 3**).

The cortical layer lies beneath the cuticle. Its long spindle cells contribute tensile strength and elasticity to the wool fiber. The cortex is further divided bilaterally into a paracortex and an orthocortex. This is demonstrated by the use of a preferential dye (acid or basic), which is taken up by the orthocortex, and may be seen in a cross-sectional view.

In medium and coarse wools, the third layer, the medulla, comprises superimposed, honeycomblike cells filled with air. Medullation is a problem to the manufacturer inasmuch as fibers possessing it have lower spinning properties and are lustrous, straight, and coarse. In piece-dyed fabrics, they produce a skittery effect if they are dyed a lighter shade.

Physical properties. Wool's major physical characteristics include fiber diameter (fineness or grade), staple length, and clean wool yield. Also significant are soundness, color, luster, and content of vegetable matter. Grade refers specifically to mean fiber diameter and its variability. Fiber diameter is the most important manufacturing characteristic. Fleeces are commercially graded visually through observation and handling by persons of long experience in the industry. Degree of crimp and relative softness of the fleece are important deciding factors employed by the graders (**Fig. 4**).

Grading systems. Two systems of grading fleeces are practiced, the American and the spinning count. The seven grades of the American system are inadequate for modern manufacturing systems. The spinning-count system (**Table 2**) provides for 14 grades, the range in fiber diameter for the various spinning counts proposed by the American Society for Testing and Materials (ASTM).

Measuring devices. For meticulous laboratory research, fiber diameter is expressed in micrometers (μm) according to methods set forth by the ASTM. Devices utilized in such work include the Hardy sectioning device (**Fig. 5**) and the micronaire (**Fig. 6**).

Advantages of grading. The accurate grading of wool is important to producer and manufacturer for the following reasons: (1) it is mechanically impossible to make soft, full-handling worsted fabric from coarse wool; (2) manufacturers must use graded wools for

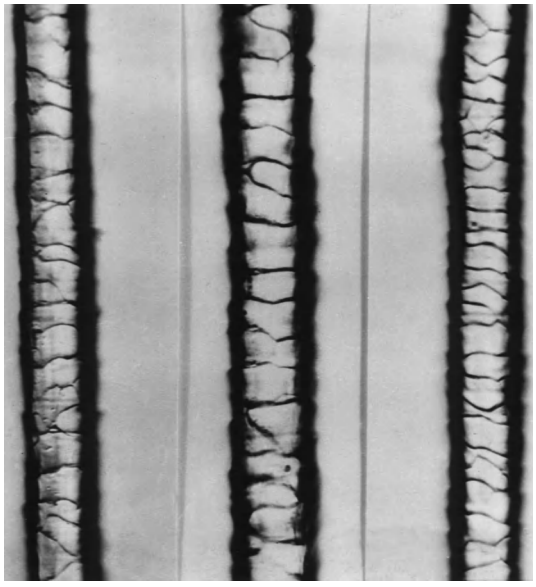


Fig. 3. Scale formation of Delaine Merino wool fibers.

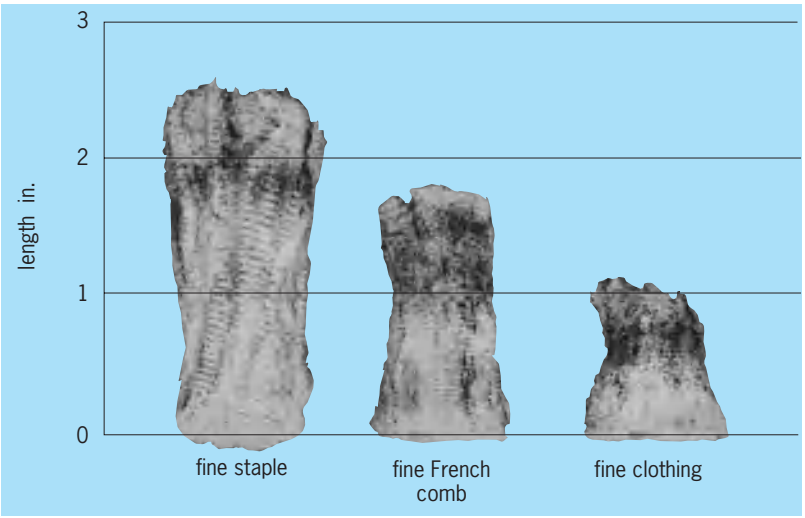


Fig. 4. Fine wool staples showing distinct crimp (wavy) pattern. The length shown for each classification is the minimum for that classification. 1 in. = 2.5 cm.

TABLE 2. Proposed specifications for grade or fineness of wool (ASTM)*

	Fineness range, μm (avg. diam.)			Fineness range, μm (avg. diam.)	
Grade	Min	Max	Grade	Min	Max
80s	17.7	19.1	54s	27.9	29.3
70s	19.2	20.5	50s	29.4	30.9
64s	20.6	22.0	48s	31.0	32.6
62s	22.1	23.4	46s	32.7	34.3
60s	23.5	24.9	44s	34.4	36.1
58s	25.0	26.4	40s	36.2	38.0
56s	26.5	27.8	36s	38.1	40.2

*Numerical terms for grade are used internationally and represent the maximum spinning capacity of wool of that fineness.

securing desired effects in finished goods; (3) textile machinery either is designed to handle certain grades exclusively, or must be adjusted for each grade; (4) graded wools gain a market advantage; (5) ranch-graded wools give the producer an index of the variability of the flock so that corrective measures can be applied in the selecting of breeding sheep; and (6) shrinkage estimates are more accurate on graded lines of wool than on ungraded lines.

Length. While wool is being graded, it is classified by length into three major categories: staple, French-combing, and clothing (**Table 3**). Usually longer wools within grade are more valuable than shorter wools because they shrink less and are less wasteful in manufacture. As fiber diameter increases, staple length increases. The length for staple classification also increases for each successively coarser grade.

Clean wool yield. Extraneous matter in unclean wool is extremely variable in quantity. Such matter consists of grease, suint, dirt, vegetable matter, and moisture. All of these except moisture are removed in the scouring (washing) process. The loss in weight may vary from 32 to 78%, and is called shrinkage; the clean scoured wool remaining is the yield.

Because wool is unmanageable after washing, the fiber is dipped in, or sprayed with, a light emulsion of olive or mineral oil to prevent it from becoming brittle and to lubricate it for the spinning operation. If the wool is to be dyed in the raw stock, it is dyed at this stage.

The percentage of extraneous matter in wool is a major factor in determining its market value. Estimates of the yield are determined either by laboratory testing of samples, or through visual appraisal by the buyer. Because of the variability in shrinkage according to length and grade, visual appraisal of clean wool yield may be inaccurate. The core test is a more reliable and consistent method.

Core test. By means of a motor-driven pressure tube, a core sample of wool is withdrawn from a bag

Fig. 5. Hardy sectioning device used in laboratory research for measuring wool fiber diameter.

Fig. 6. The micronaire is a device for measuring wool fiber diameter, which is expressed in microns (micrometers).

TABLE 3. Commercial lengths and grades of wool

American system:	Fine	1/2 Blood	3/8 Blood	1/4 Blood	Low 1/4 blood	Common	Braid
Spinning count:	80s–70s, 64s	62s–60s	58s–56	54s–50s	48s–46s	44s	40s–36s
Commercial length classes	**Staple length by grade in inches***						
Staple	2.5 and longer	3.0 and longer	3.5 and longer	4.0 and longer	4.5 and longer	5.0 and longer	5.5 and longer
Good French-combing	2.0	2.5	3.0	3.5			
Average French-combing	1.5	2.0	2.0	2.5			
Short French-combing	1.0	1.5					
Clothing and stubby	Under 1.0	Under 1.5	Under 2.0	Under 2.5	Under 4.5	Under 5.0	Under 5.5

*The length designations are based on unstretched staple length and represent a minimum length for the bulk of the staples in a sample. 1 in. = 2.5 cm.

or bale. Core-sampling patterns and the number of cores taken are prescribed according to the number of bags or bales in the lot. Each sample is then tested by standard procedures for moisture, ash, vegetable matter, and grease content, and is then adjusted to the standard condition of 14% impurities or 12.0% moisture, 0.5% ash, and 1.5% residual grease.

Chemical characteristics. Wool is primarily a protein, keratin (**Table 4**). Its molecular structure consists of long polypeptide chains in which 17 or more known amino acids are linked together by the disulfide groups of cystine. Reagents which alter the disulfide linkages change the physical characteristics of the fiber as a whole. Such reagents are oxidizers, reducers, alkalies, and light. The amphoteric nature of wool is a ready-made tool for the dyer and colorist.

Action of halogens. Treatment of the fiber with the halogens leads to absorption and chemical change. Chlorination causes wool to become yellow, harsh, and lustrous, and to lose its felting characteristics, with a corresponding rise in rate of dye absorption.

Action of heat. If briefly heated in dry air at 100–105°F (37–40°C), wool becomes harsh and loses strength and moisture. Normal moisture, softness, and strength are regained upon return to moist, cool conditions. Wool decomposes over extended periods of heat treatment.

Action of cold. At subzero temperatures, wool remains pliable and undergoes no perceptible chemical change.

Action of water and steam. Normally, wool is insoluble in water, though highly hygroscopic. If wool is boiled for 2 h, a weight loss of about 25% occurs. Conversely, the fiber diameter will swell approximately 10% with no damaging effects during brief periods of steaming (212°F or 100°C). Wool is more efficiently converted into yarn when it has been kept in a warm, moist environment; however, prolonged steaming causes loss of strength. Wool is not allowed to become absolutely dry. Usually 12–16% of the moisture

TABLE 4. Amino acid composition of wool*

Amino acid	Approximate percentage present in wool	Percentage of residue by weight	Percentage of side chain by weight
Glycine	6.5	4.49	0.09
Alanine	4.4	3.52	0.74
Serine	9.41	7.80	2.76
Proline	6.75	5.69	2.46
Valine	4.72	3.99	1.73
Threonine	6.76	5.74	2.59
Cystine	12.72†	10.83	4.89
Leucine isomers	11.3	9.75	4.92
Aspartic acid	7.27	6.28	3.22
Lysine	3.3	2.89	1.63
Glutamic acid	15.27	13.40	7.58
Methionine	0.71	0.62	0.36
Histidine	0.7	0.62	0.37
Hydroxylysine	0.21	0.19	0.11
Phenylalanine	3.75	3.34	2.07
Arginine	10.4	9.33	5.97
Tyrosine	5.8	5.23	3.43
Tryptophan	0.7	0.64	0.45
Total	110.67	94.80	45.37
Ammonia nitrogen	1.18	−0.30	−0.30
Total, corrected for ammonia nitrogen		94.50	45.07

*After J. M. Matthews.
†Based on 3.55% total sulfur and subtracting methionine sulfur.

is left in the wool to condition it for subsequent handling.

Plasticity. Under moist conditions wool becomes plastic. Its shape can be altered and its affinity for dye changed. Knowledge of this characteristic enables the manufacturer to set yarns and to produce desired color effects. Plasticity increases with rising temperatures to a point at which stretched fibers become permanently set and will not return to their normal state.

Technological advances. The inherent advantages of wool have been exploited, and its limitations as a textile fiber have been overcome by the application of technology to manufacturing processes. The use of the insecticide dieldrin as a dye renders wool mothproof for life. Permanent pleats have been imparted to garments which are shrinkproofed and can be home laundered. Each such technological advance enables wool to hold its competitive place in the field of textile manufacture and use. *See* TEXTILE.

Thomas D. Watkins

Reuse of wool fibers. There has never been a sufficient supply of new wool stocks to take care of a steadily increasing demand for wool. To meet this situation, wool fibers have had to be recovered from old clothing, rags of all kinds, and waste from wool manufacturing. This wool is variously called salvaged, reclaimed, reworked, or remanufactured, but it is best known in the textile industry as shoddy. This term is misunderstood by the average consumer, who is inclined to believe that wool fabric containing remanufactured fibers is necessarily of inferior quality.

The hardier, although less resilient, remanufactured fibers, when obtained from good original stock and combined with new wool from sheep, add durability to the soft new wool. Thus, remanufactured fibers contribute ability to withstand hard wear, although there is some sacrifice in warmth, softness of texture, and resiliency. They also make wool clothing available at lower prices.

To correct wrong impressions concerning the use of remanufactured wool, and also to protect consumers against unscrupulous practices, the United States government passed the Wool Products Labeling Act. This provides that every article of wool clothing must be labeled according to the type of wool used in its manufacture. The label must state (1) amount of wool fiber in the fabric; (2) percentage by weight of new or virgin wool fibers; (3) percentages of reprocessed or reused fibers; (4) percentage of each fiber other than wool, if such fibers constitute 5% or more of the total; (5) aggregate of other fibers; and (6) nonfibrous loading, filling, or adulterating substance.

The term wool, according to the United States government standards, must always mean new wool, not made up from any form of wool product. New wool comes directly from a fleece. It has never been previously spun, woven, felted, or worn.

Reprocessed wool. According to the government classification, reprocessed wool is that which has been reclaimed and remanufactured from unused wool materials. Such materials may be combings and scraps of wool obtained during the manufacturing processes, sample swatches, or pieces of all-wool cloth from apparel manufacturing.

Reused wool. The United States government gives the special classification of reused wool to fiber that is salvaged from all kinds of used consumers' goods.

Virgin wool. This term is now used by the textile industry to designate new wool from a lamb's fleece, but the term is too all-inclusive to serve as a criterion of quality. Although the term testifies to the fact that virgin wool does not contain remanufactured-wool fibers, it does not distinguish between the less desirable fibers of a fleece and a specially fine quality of wool. Virgin wool may also include pulled or dead wool, which may be of definitely inferior quality. One should not feel that a fabric labeled "100% new wool" is necessarily more serviceable than one containing any of the remanufactured wool fibers.

Wool of different grades may be blended or mixed together. It is not uncommon for inferior grades to be mixed with the better grades. The use of a mixture with a coarser grade of fiber is a legitimate practice if the purpose is to make a better-wearing and less expensive product, provided the label on the finished goods indicates a true description of the raw materials used. *See* ALPACA; CAMEL'S HAIR; CASHMERE; LLAMA; MOHAIR; SHEEP; VICUNA.

M. David Potter

Word processing

The use of a computer and specialized software to write, edit, format, print, and save text. In addition to these basic capabilities, the latest word processors enable users to perform a variety of advanced functions. Although the advanced features vary among the many word processing applications, most of the latest software facilitates the exchange of information between different computer applications, allows easy access to the World Wide Web for page editing and linking, and enables groups of writers to work together on a common project. The latest word processors continue to provide increased features with each new revision. *See* COMPUTER; SOFTWARE; WORLD WIDE WEB.

Writing. Writing is accomplished by using the computer's typewriterlike keyboard. Most of the characters shown on the keyboard can be typed. The characters appear on the computer screen as they are typed. A finite number of characters can be typed across the computer screen. The word processor "knows" when the user has reached this limit and automatically moves the cursor to the next line for uninterrupted typing.

The position on the computer screen where a character can be typed is marked by a blinking cursor, which is most commonly a vertical bar or underline character. The cursor can be positioned anywhere on the screen by using the mouse, or the keys marked with arrows on the keyboard. When using the arrow keys, the cursor moves in the direction

shown on the key. However, if there are no characters typed on the computer screen, the cursor will not move. *See* COMPUTER PERIPHERAL DEVICES; ELECTRONIC DISPLAY.

In addition to writing, the latest word processors provide tools to create and insert drawings anywhere in the document. Typical features allow users to draw lines, rectangles, circles, and arrowheads, and to add text. Some even enable the creation of three-dimensional effects. A variety of color options are available for lines and text, and users can also fill or shade an object, such as a circle or rectangle, with a specified color.

Editing. Editing is the process of changing an existing document. Users can, correct typographical errors, add new sentences or paragraphs, move entire blocks of text to a different location, delete portions of the document, copy text and paste it somewhere else in the document, or insert text or graphics from an entirely different document.

Most word processing programs can automatically correct many basic typographical errors, such as misspelled words, two successive capital letters in a word, and failure to capitalize the first letter of the names of days and of the first word in a sentence. Some other helpful editing tools commonly found in word processors include an automatic spelling checker, a thesaurus, and a grammar checker.

Formatting. Formatting enables users to define the appearance of the elements in a document, such as the font and type size of all headings and text, the left, right, top, and bottom margins of each page, and the space before and after sentences and paragraphs. Most word processors allow all the elements in a document to be formatted at once. This is accomplished by applying a "style." Typically, a style is selected from a library of styles provided with the word processor. There are styles for many different types of documents, including letters, resumes, and brochures. Once an overall style has been applied to a document, the style characteristics of any element within the document can be changed.

Word processors are approaching the formatting power of full-featured desktop publishing applications. The formatted page can be viewed on the computer screen exactly as it will be printed. This is referred to as "what you see is what you get" (WYSIWYG). In addition, users can add text and graphics anywhere on a page, and blocks of text and graphics can be uniquely formatted on the same page. For instance, one block of text can be in single-column format, while another block of text on the same page is in multiple-column format.

Printing. The most common printing operation is to send the document electronically from the word processor to a printer in order to produce a paper copy. The quality of the printed document is a function of the printer and, to a lesser degree, the paper, and can vary significantly. The output quality, measured in dots per inch (dpi), is directly proportional to the quality of the ink. The higher the dpi, the sharper the printed image and the more time required to print.

Other common print features include printing to a file, printing to a fax machine, and previewing pages before printing.

Printing to a file. Word processors allow a document to be printed to a file rather than to a printer. This feature is commonly used to send files directly to a print shop. Many print shops have printing equipment that can read these files directly and produce extremely high-quality printed output. *See* DATA COMMUNICATIONS.

Printing to a fax machine. Another common print feature is the ability to send an electronic document to a fax machine, which produces a much higher-quality fax than sending paper copies from one fax machine to another. *See* FACSIMILE.

Print previewing. Print previewing allows users to view the pages of a document before printing. This option displays the pages exactly as they will be printed, which allows for further editing, if required, and also saves paper.

Saving. Saving is the process of storing a copy of the electronic document as a file on a floppy or compact disk or the computer's hard disk drive. Word processors can be configured to automatically perform a save operation while working on the document, which helps prevent the loss of work.

The "Save As" feature is used to store a copy of the document under a different name or to save the file to a new location on the disk, a network node, or some other storage portable media such as a diskette or Zip disk. *See* COMPUTER STORAGE TECHNOLOGY.

Exchanging information between applications. Many business and home users purchase word processors as part of a group of applications. The other applications are typically spreadsheet, database, presentation, and personal information management applications. This package of applications is referred to as a software suite. Software suites allow for easy exchange of information. Portions of a spreadsheet, for instance, can be copied and pasted directly into a word processing document. Entire files can even be inserted. In this way, information such as text files, drawings, pictures, and spreadsheets, can be combined into a single document.

This exchange capability is made possible by a function referred to as object linking and embedding (OLE). Linking refers to the ability to access information from more than one application simultaneously; embedding refers to the ability to incorporate a copy of information from one application into another. Embedded information is referred to as an object. In order to link and embed information, both applications must support OLE.

Creating and editing Web pages. Word processing programs can be used to create and edit Web pages for the Internet. Some programs have built-in features for guiding a novice Web page designer through the entire creation process. Once created, Web pages can be edited and saved as easily as any file on a computer's hard drive. Links can be inserted into a Web page that can access other Web addresses. Also, Web pages can be viewed within the word processor, so an Internet browser is not necessary for checking

edits that have been made. In addition, existing word processing documents can easily be converted into Web pages. *See* INTERNET; WORLD WIDE WEB.

Working in groups. The latest word processors have many features for allowing groups of people to work together on the same document. For instance, multiple versions of a document can be saved to a single file for version control; access levels can be assigned so that only a select group of people can make changes to a document; edits can be marked with the date, time, and editor's name; and text colors can be assigned to differentiate editors. In addition, some word processors have editing features that include highlighting text, drawing lines through text to represent deleted text, and using red underscoring to identify changed text.

Carlos Quiroga

Bibliography. A. Diller, *LaTeX Line by Line: Tips and Techniques for Document Processing*, 2d ed., John Wiley, 1999; R. Person, *Using Microsoft Word 97*, Que, 1997; R. Williams, *The PC Is Not a Typewriter: A Style Manual for Creating Professional-Level Type on Your Personal Computer*, Peachpit Press, 1992.

Work

In physics, the term work refers to the transference of energy that occurs when a force is applied to a body that is moving in such a way that the force has a component in the direction of the body's motion. Therefore work is done on a weight that is being lifted, or on a spring that is being stretched or compressed, or on a gas that is undergoing compression in a cylinder.

When the force acting on a moving body is constant in magnitude and direction, the amount of work done is defined as the product of just two factors: the component of the force in the direction of motion, and the distance moved by the point of application of the force. Thus the defining equation for work W is (1), where f and s are the magnitudes of the force

$$W = f \cos \phi \cdot s \qquad (1)$$

and displacement, respectively, and ϕ is the angle between these two vector quantities (**Fig. 1**). Because $f \cos \phi \cdot s = f \cdot s \cos \phi$, work may be defined alternatively as the product of the force and the component of the displacement in the direction of the force. In

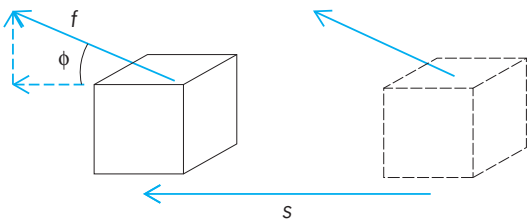


Fig. 1. Work done by constant force *f* is *fs* cos ϕ.

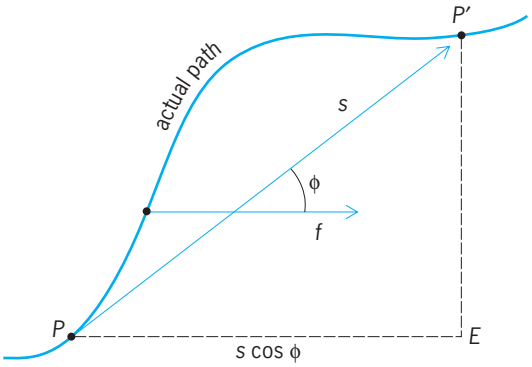


Fig. 2. Work done in traversing any path connecting points *P* and *P′* is *f* · $\overline{PE}$, assuming the force *f* to be constant in magnitude and direction.

Fig. 2 the work done by the constant force f when the application point moves along the curved path from P to P', and therefore undergoes the displacement $\overline{PP'}$, is $f \cdot \overline{PP'} \cos \phi$, or $\overline{PE}$.

Work is a scalar quantity. Consequently, to find the total work on a moving body by several different forces, the work of each may be computed separately and the ordinary algebraic sum taken.

Examples and sign conventions. Suppose that a car slowly rolls forward a distance of 10 m along a straight driveway while a man pushes on it with a constant magnitude of 200 newtons of force (200 N) and let Eq. (1) be used to compute the work W done under each of the following circumstances: (1) If the man pushes straight forward, in the direction of the car's displacement, then $\phi = 0^\circ$, $\cos \phi = 1$, and $W = 200 \text{ N} \times 1 \times 10 \text{ m} = 2000 \text{ N} \cdot \text{m} = 2000$ joules; (2) if he pushes in a sideways direction making an angle ϕ of 60° with the displacement, then $\cos 60^\circ = 0.50$ and $W = 1000$ joules; (3) if he pushes against the side of the car and therefore at right angles to the displacement, $\phi = 90^\circ$, $\cos \phi = 0$, and $W = 0$; (4) if he pushes or pulls backward, in the direction opposite to the car's displacement, $\phi = 180^\circ$, $\cos \phi = -1$, and $W = -2000$ J.

Notice that the work done is positive in sign whenever the force or any component of it is in the same direction as the displacement; one then says that work is being done *by* the agent exerting the force (in the example, the man) and *on* the moving body (the car). The work is said to be negative whenever the direction of the force or force component is opposite to that of the displacement; then work is said to be done *on* the agent (the man) and *by* the moving body (the car). From the point of view of energy, an agent doing positive work is losing energy to the body on which the work is done, and one doing negative work is gaining energy from that body.

Units of work and energy. These consist of the product of any force unit and any distance unit. Units in common use are the foot-pound, the foot-poundal, the erg, and the joule. The product of any power unit and any time unit is also a unit of work or energy. Thus the horsepower-hour (hp-h) is equivalent, in view of the definition of the

horsepower, to 550 ft-lbf/s × 3600 s, or 1,980,000 ft-lbf, or (1,980,000)(0.3048 m)(4.45 N) = 2,684,520 J. Similarly, the watt-hour is 1 J/s × 3600 s, or 3600 J; and the kilowatt-hour is 3,600,000 J.

Work of a torque. When a body which is mounted on a fixed axis is acted upon by a constant torque of magnitude τ and turns through an angle θ (radians), the work done by the torque is $\tau\theta$.

Work principle. This principle, which is a generalization from experiments on many types of machines, asserts that, during any given time, the work of the forces applied to the machine is equal to the work of the forces resisting the motion of the machine, whether these resisting forces arise from gravity, friction, molecular interactions, or inertia. When the resisting force is gravity, the work of this force is mgh, where mg is the weight of the body and h is the vertical distance through which the body's center of gravity is raised. Note that if a body is moving in a horizontal direction, h is zero and no work is done by or against the gravitational force of the Earth. If a person holds an object or carries it across level ground, she does no net work against gravity; yet she becomes fatigued because her tensed muscles continually contract and relax in minute motions, and in walking she alternately raises and lowers the object and herself.

The resisting force may be due to molecular forces, as when a coiled elastic spring is being compressed or stretched. From Hooke's law, the average resisting force in the spring is $-\frac{1}{2}ks$, where k is the force constant of the spring and s is the displacement of the end of the spring from its normal position; hence the work of this elastic force is $-\frac{1}{2}ks^2$. *See* HOOKE'S LAW.

If a machine has any part of mass m that is undergoing an acceleration of magnitude a, the resisting force $-ma$ which the part offers because of its inertia involves work that must be taken into account; the same principle applies to the resisting torque $-I\alpha$ if any rotating part of moment of inertia I undergoes an angular acceleration α.

When the resisting force arises from friction between solid surfaces, the work of the frictional force is $\mu f_n s$, where μ is the coefficient of friction for the pair of surfaces, f_n is the normal force for the pair of surfaces together, and s is the displacement of the one surface relative to the other during the time under consideration. The frictional force μf_n and the displacement s giving rise to it are always opposite in direction ($\phi = 180°$). *See* FRICTION.

The work done by any conservative force, such as a gravitational, elastic, or electrostatic force, during a displacement of a body from one point to another has the important property of being path-independent: its value depends only on the initial and final positions of the body, not upon the path traversed between these two positions. On the other hand, the work done by any nonconservative force, such as friction due to air, depends on the path followed and not alone on the initial and final positions, for the direction of such a force varies with the path,

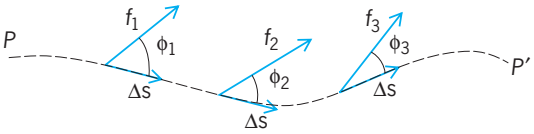


Fig. 3. Work done by a variable force.

being at every point of the path tangential to it. *See* FORCE.

Since work is a measure of energy transfer, it can be calculated from gains and losses of energy. It is useful, however, to define work in terms of forces and distances or torques and angles because these quantities are often easier to measure than energy changes, especially if energy changes are produced by nonconservative forces.

Work of a variable force. If the force varies in magnitude and direction along the path $\overline{PP'}$ of its point of application, one must first divide the whole path into parts of length Δs, each so short that the force component $f\cos\phi$ may be regarded as constant while the point of application traverses it (**Fig. 3**). Equation (1) can then be applied to each small part and the resulting increments of work added to find the total work done. Various devices are available for measuring the force component as a function of position along the path. Then a work diagram can be plotted (**Fig. 4**). The total work done between positions s_1 and s_2 is represented by the area under the resulting curve between s_1 and s_2 and can be computed by measuring this area, due allowance being made for the scale in which the diagram is drawn.

For an infinitely small displacement ds of the point of application of the force, the increment of work dW is given by Eq. (2), a differential expression that

$$dW = f\cos\phi\, ds \tag{2}$$

provides the most general definition of the concept of work. In the language of vector analysis, dW is the scalar product of the vector quantities $\mathbf{f}$ and ds; Eq. (2) then takes the form $dW = \mathbf{f}\cdot ds$. If the force is a known continuous function of the displacement, the total work done in a finite displacement from point P to point P' of the path is obtained by evaluating the

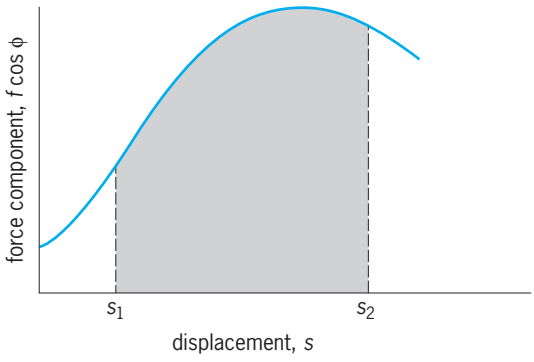


Fig. 4. Work diagram.

line integral in Eq. (3).

$$W = \int_P^{P'} f \cos\phi \, ds = \int_P^{P'} \mathbf{f} \cdot ds \qquad (3)$$

When a variable torque of magnitude τ acts on a body mounted on a fixed axis, the work done is given by

$$W = \int_{\theta_1}^{\theta_2} \tau \, d\theta$$

where $\theta_2 - \theta_1$ is the total angular displacement expressed in radians. *See* ENERGY. Leo Nedelsky

Bibliography. F. Bueche, *Principles of Physics*, 6th ed., 1994; D. Halliday and R. Resnick, *Fundamentals of Physics*, 6th ed., 2000; H. C. Ohanian, *The Principles of Physics*, 1994.

Work function (electronics)

A quantity with the dimensions of energy which determines the thermionic emission of a solid at a given temperature. The thermionic electron current density J emitted by the surface of a hot conductor at a temperature T is given by the Richardson-Dushman formula, $J = AT^2e^{-\phi/kT}$, where A is a constant, k is Boltzmann's constant ($= 1.38 \times 10^{-23}$ joule per degree Celsius) and ϕ is the work function; the last may be determined from a plot of $\log (J/T^2)$ versus $1/T$. For metals, ϕ may also be determined by measuring the photoemission as a function of the frequency of the incident electromagnetic radiation; ϕ is then equal to the minimum (threshold) frequency for which electron emission is observed times Planck's constant h ($= 6.63 \times 10^{-}$ joule second). The work function of a solid is usually expressed in electronvolts (1 eV is the energy gained by an electron as it passes through a potential difference of 1 V, and is equal to 1.60×10^{-19} J). A list of average values of work functions (in electronvolts) for metals is given in the **table**.

The work function of metals varies from one crystal plane to another and also varies slightly with temperature (approximately 10^{-4} eV/degree). For a metal, the work function has a simple interpretation. At absolute zero, the energy of the most energetic electrons in a metal is referred to as the Fermi energy; the work function of a metal is then equal to the energy required to raise an electron with the Fermi energy to the energy level corresponding to an electron at rest in vacuum. The work function of a semiconductor or an insulator has the same interpretation, but in these materials the Fermi level is in general not occupied by electrons and thus has a more abstract meaning. *See* FIELD EMISSION; PHOTOEMISSION; THERMIONIC EMISSION. Adrianus J. Dekker

Average values of work functions for metals, in electronvolts

Metal	Value	Metal	Value	Metal	Value
Al	4.20	Cs	1.93	Na	2.28
Ag	4.46	Cu	4.45	Ni	4.96
Au	4.89	Fe	4.44	Pd	4.98
Ba	2.51	K	2.22	Pt	5.36
Cd	4.10	Li	2.48	Ta	4.13
Co	4.41	Mg	3.67	W	4.54
Cr	4.60	Mo	4.24	Zn	4.29

Work function (thermodynamics)

The thermodynamic function better known as the Helmholtz energy, $A = U - TS$, where U is the internal energy, T is the thermodynamic (absolute) temperature, and S is the entropy of the system. At constant temperature, the change in work function is equal to the maximum work that can be done by a system ($\Delta A = w_{max}$). *See* FREE ENERGY. P. W. Atkins

Work measurement

The determination of a set of parameters associated with a task.

Rationale. There are four reasons, common to most organizations whether profit seeking or not, why time, effort, and money are spent to measure the amount of time a job takes. The fifth, pay by results, is used only by a minority of organizations.

Cost accounting. If the length of time it takes to do the job is not known, the charge for the job cannot be determined. Without knowing the time/job ratio, too much or too little may be charged. As a result, the product may be overpriced and, in the long run, sales will be lost, or if too little is charged, revenue will be lost.

Evaluation of alternatives. Without knowing the time (and thus the cost) of a job, decisions regarding whether to make or buy an item, whether to mechanize or not, whether to advertise or not, could be wrong.

Acceptable day's work. Suppose a worker makes 100 widgets a day. A standard of comparison is needed to determine whether this is superior, average, or poor performance. If the standard is 75 a day, the worker is to be praised; if the standard is 200 a day, the worker is to be questioned.

Scheduling. Managers need to know the time/job ratio to make reasonable decisions for production schedules, how many people to assign to the job, how much equipment is required, and so on.

Pay by results. In a small number of organizations, pay is based on the units produced. In most organizations, pay is by the hour, week, or month and, as long as output is "reasonable," pay continues. But if pay depends upon units produced, the organization needs to know how long the task takes so a reasonable pay rate/unit can be set. *See* WAGE INCENTIVES.

Given the decision to record the amount of time per job, it is important to emphasize that first the job should be properly designed. Recording time for a poorly designed job is a mistake. The job should be improved (that is, improve productivity) before doing work measurement. *See* PRODUCTIVITY.

There are three common ways to determine time

NO.	ELEMENTS	SPEED	FEED	UPPER LINE: SUBTRACTED TIME 1	2	3	4	5	6	7	8	LOWER LINE: READING 9	10	11	12	13	14	15	MIN. TIME	AV. TIME	Std. Dev.	OCC. PER CYCLE	EFFORT RATING	NORMAL TIME
1	Get pen; TP Ready to use			3.1																3.1		.25	100	.775
2	Get letter & envelope: TP Ready to check			2.6	2.7	3.5	3.0													2.95		1	90	2.655
3	Read letter: TP Read last letter: TP Read last word *			18.2 (100)	21.5 (125)	42.1 (225)	55.0 Ⓐ (205)					18.2/100 = .182/word 21.5/125 = .172 42.1/225 = .187								.180/ word			100	.180/ word
4	Sign and aside letter and envelope: TP RL paper			3.0	2.7	3.1	3.0																	
5	Return pen: TP RL pen						3.6													3.6		.25	110	.990
	* Number of words in bracket																							
																						$\frac{4.42}{9} = 4.91$		
																						$\frac{.180}{9} = .2$/word		

FOREIGN ELEMENTS:

TOOLS, JIGS, GAUGES, PATTERNS, ETC.:

Ⓐ Made 1 correction

OVERALL EFFORT RATING	BEGIN	END	ELAPSED	UNITS FINISHED	ACTUAL TIME PER PIECE

Elemental breakdown of a job: signing a letter. The end of the element is the termination point (TP). The times are seconds. It is assumed that four letters are considered for signature at a time, and a standard will assume no changes are needed in the letter. Elements 1, 2, 4, and 5 are constant elements. Element 3 is a variable element as the time varies with the number of words read. Allowances are assumed as 10%.

per job: stopwatch time study (sequential observations), occurrence sampling (nonsequential observations), and standard data.

Stopwatch time study. Stopwatch time study has the great advantage of flexibility—it can be used for almost any existing job. Workers have become familiar with it. It is reasonable in cost and gives reasonable accuracy. However, it does require the worker to be rated. Once the initial cost of a standard data system has been incurred, standard data may be the lowest-cost, most accurate, and most accepted technique.

There are eight steps to determine standard time:

1. *Development of a good method to be timed.*

2. *Selection of an operator to time.* In some cases, there is no choice as only one person does the job. If there is a choice, a typical or average worker should be used rather than the best. (The best is useful for productivity analysis.) An average worker is used because rating is more accurate for that worker than for an extreme worker and because other workers are more willing to accept a time standard determined on an average worker.

3. *Preparation for timing.* The job must be broken down into elements, indicating the termination points (see **illus.**). Elements permit reuses of the data, they give good internal consistency checks, they permit different ratings for different elements, and they improve methods descriptions.

4. *Selection of a timing technique.* There are four common alternatives: one watch with continuous hand movement; one watch with snapback; three watches with snapback; and an electronic watch with a hold circuit. The one watch with continuous hand movement formerly was the recommended procedure but has been made obsolete by the improved technology and lower costs of the third and fourth methods. While the first method gives accurate results, it requires excessive clerical work. The last two methods give even better accuracy and substantially reduced clerical work, at a minor incremental equipment cost for the watch. The second method has low clerical costs, but has considerable potential for inaccuracy.

5. *Determination of the number of observations.* The number can be determined by using statistical techniques from standard tables used by the organization. For example, the organization may require 200 observations for any element time less than .002 h, 175 for times from .002 to .003, and so on.

6. *Elimination of foreign elements, irregular elements, and outliers from the data.* Foreign elements are observations which are not allowed directly as part of the time standard—they may be included in

some cases indirectly as allowances. Examples of foreign elements would be speaking to a supervisor, watching someone walk by, and lighting a cigarette. An irregular element (such as oiling of the machine) occurs at infrequent intervals. The data should be used, but the number of units between elements must be determined. An outlier is an abnormally high or low value; it can be eliminated only for specific known reasons or with standard statistical tests—never merely because it is high or low.

7. *Performance rating.* Completion of the first six steps gives the recorded time. The normal time is needed—the time that a typical experienced worker would take under ideal conditions and without a break: Recorded time × rating = normal time.

8. *Allowances.* Allowances generally fall into personal, fatigue, and delay allowances. Personal allowances would be for getting a drink of water; fatigue allowances are given for physically or mentally demanding work; and delay allowances are given for uncontrollable delays. Total allowances generally range from 5 to 20% of the standard time for the job or operation. Standard time is calculated from normal time divided by (1 − allowance percent).

Occurrence sampling standards. Occurrence sampling is also called work sampling or ratio-delay sampling. If time study is a "movie," then occurrence sampling is a "series of snapshots."

The primary advantage of this approach may be that occurrence sampling standards are obtained from data gathered over a relatively long time period, so the sample is likely to be representative of the universe. That is, the time from the study is likely to be representative of the long-run performance of the worker. Another advantage is that occurrence sampling can be used when "production" is not continuous. For example, when timing phone calls from customers, calls may occur only 5 to 10 times per day and at irregular intervals. Rating generally is not used as all work is assumed to be at a 100% pace. For work done at paces different from 100%, this will cause errors in the accuracy of the standard. Thus occurrence sampling standards probably should not be used for incentive pay purposes.

To set a standard using occurrence sampling, it is necessary to do an occurrence sample plus record the units produced during the time the occurrence sample takes place. For example, consider the mechanic who does tune-ups. The individual should be observed 100 times over a 10-day period, with records of each period of idle time and of each truck or car tune-up. The output during the 10 days (say 5 trucks and 13 cars) is recorded as well as the scheduled work time (say 450 min/day).

Assume the study showed idle = 10%, truck tune-up = 36%, and car tune-up = 45%. Then working time was .9(4500) = 4050 min; a truck tune-up takes .36(4050)/5 = 292 min, and a car tune-up takes 168/.9 = 187 min.

In making the occurrence sample there are two steps: getting a sample whose size gives the desired trade-off between cost of the study and risk of an inaccurate estimate, and obtaining a sample representative of the population.

TABLE 1. Confidence levels for z levels on occurrence sampling formula*

z (number of standard deviations)	Corresponding confidence level, %
±1.0	68
±1.64	90
±1.96	95
±2.0	95.45
±3.0	99.73

*Use the table of the normal distribution (not shown here) for other values.
SOURCE: S. A. Konz, *Work Design*, published by Grid, Columbus, Ohio, 1979.

The required number of observations in the sample can be determined from the formula below,

$$A = z\sigma_p$$

where $A = sp$ = absolute accuracy desired, decimal; s = relative accuracy desired, decimal; p = mean percent occurrence, decimal; z = number of standard deviations for confidence level desired (**Table 1**); n = number of observations; and σ_p = standard deviation of a percent, $\sqrt{p(1-p)/n}$.

For example, the problem might be to determine whether or not to add an additional telephone line, as customers have been complaining the line is always busy. Management requests a study, indicating that they want a relative accuracy of ±10% and a confidence of 90%. The person doing the study must decide how many observations (samples) to take. First, from preliminary judgment, it is guessed that the lines are busy 60% of the time—that is, $p = .60$. Then $A = sp$ or .10 (.60) = .06. The number of standard deviations corresponding to a 90% confidence level is 1.64. The estimated value of p (.6) is then substituted into the formula for standard deviation of a percent. The resulting equation is $.06 = (1.64)\sqrt{(.6)(.4)/n}$. Solving for n, which is 179, and assuming 2 weeks would be needed to give a representative sample, 180 observations must be taken at the rate of 18/day for 10 days.

To make the sample representative, stratification, influence, and periodicity must be considered. Stratification means to divide the sample into strata (layers). Thus there may be 10 strata (5 days/week for 2 weeks), 20 strata (morning versus afternoon for each day of the sample), or some other division such as local calls versus long-distance. Influence means that it is not desirable that the behavior of the individual changes because of being observed. For this example, observations should be taken in such a way that neither customers nor employee change their use of the telephone because they are being sampled. Periodicity refers to existing patterns of behavior coincident with specific times of the day. The sample must not take too many or too few observations at these special times.

Assume that from the sample of 180 the lines were busy 90 times, so $p = 90/180 = .5$. Therefore,

$A = 1.64\sqrt{(.5)(.5)/180}$ so $A = 6.1\%$. Thus the conclusion is that the phones are busy 50% (±6.1%) of the time with a confidence of 90%—if the situation does not change. The stratification information indicates that the percent busy was 40% on Monday and Wednesday, 50% on Tuesday, and 60% on Thursday and Friday.

To set a time standard for phone calls, all that is needed is to record the number of calls during the time of the study. It is already known that the line was busy 50% of the time of 480 min/day × 10 days × .50 = 2400 min. If there were 512 calls during the 2 weeks, then time/call is 2400/512 = 4.7 min/call.

From this information, several actions are possible, such as installing another telephone line, estimating the cost of dealing with customers over the telephone, and estimating how much extra time is needed for the telephone on Thursday and Friday (and how much is available on other days). However, a new phone line will change the situation, so the times from the sample may not be representative of the situation after the change.

Standard data standards. Reuse of previous times (standard data) is an alternative to measuring new times for an operation. Lower cost, consistency, and ahead-of-production are the three advantages. Looking up the time to walk 160 ft (50 m) from a table of standard times rather than doing a special study to measure the time to walk 160 ft (50 m) saves the time of the person setting the standard. However, this operating cost must be balanced by the capital cost of setting up the table of standard times for walking. Thus the capital cost must be justified by many uses—that is, standard data are economically justifiable only for standard repetitive elements. A second advantage is consistency. In setting the time to walk 160 ft (50 m), a number of studies can be used, so any rating or measurement errors would tend to average out. In addition, since a table or formula is used, every analyst always gets the same answer; rating and judgment are minimized. A third advantage is that the time can be estimated prior to production. Timing requires an experienced operator, a work station with tools, and product. But times often are wanted ahead of production for determining which alternative work method to use and such.

There are three levels of detail: micro, elemental, and macro (**Table 2**). Microlevel systems have times of the smallest component ranging from about .01 to 1 s. Components usually come from a predetermined time system such as methods-time-measurement (MTM) or Work-Factor. Elemental level systems have the time of the smallest component, ranging from about 1 to 1000 s. Components come from time study or microlevel combinations. Macrolevel systems have times ranging upward from about 1000 s. Components come from elemental-level combinations, from time studies, and from occurrence sampling.

For example, assume a standard time is needed for signing a business letter (see illus.). It may be decided to break the task into five elements: (1) get a pen, (2) get a letter and envelope, (3) read the letter, (4) sign and set aside the letter and envelope, and (5) return pen. For simplicity it may be assumed that all letters are examined in batches of four, and no corrections need to be made. Element 3, read the letter, is a variable element in that it varies with the letter length. The remaining elements are constant. Element 3 time might be 5. 05N, where N = the number of words in the letter. If the total time for elements 1, 2, 4, and 5 was 150 time-measurement units (TMU), then total time for signing a letter is 150 + 5.05N. Thus a 100-word letter takes 150 + 505 = 655 TMU, and a 200-word letter takes 150 + 1010 = 1160 TMU. Allowances should be added to both times, but rating is not necessary because it is built into the MTM time. *See* METHODS ENGINEERING; PERFORMANCE RATING.

Stephan A. Konz

TABLE 2. Three levels of detail for standard time systems

Microsystem (typical component time range from .01 to 1 s; MTM nomenclature)

Element	*Code*	*Time*
Reach	R10C	12.9 TMU*
Grasp	G4B	9.1
Move	M10B	12.2
Position	P1SE	5.6
Release	RL1	2.0

Elemental system (typical component time range from 1 to 1000 s)

Element	*Time*
Get equipment	1.5 min
Polish shoes	3.5
Put equipment away	2.0

Macrosystem (typical component times vary upward from 1000 s)

Element	*Time*
Load truck	2.5 h
Drive truck 200 km	4.0
Unload truck	3.4

*27.8 TMU = 1 s; 1 s = .036 TMU.
SOURCE: S. A. Konz, *Work Design*, published by Grid, Columbus, Ohio, 1979.

Bibliography. S. A. Konz, *Work Design: Industrial Ergonomics*, 5th ed., 1999.

Work standardization

The establishment of uniformity of technical procedures, administrative procedures, working conditions, tools, equipment, workplace arrangements, operation and motion sequences, materials, quality requirements, and similar factors which affect the performance of work. It involves the concepts of design standardization applied to the performance of jobs or operations in industry or business. The task of work standardization thus follows the design process and precedes the establishment of systems of work measurement. The latter are designed to enforce work standards. *See* DESIGN STANDARDS; WORK MEASUREMENT.

Work standardization is a part of methods engineering and usually precedes the setting of time

standards. There would be little point in establishing time standards until the method by which the work is to be done has been stan- dardized.

The objectives of work standardizations are lower costs, greater productivity, improved quality of workmanship, greater safety, and quicker and better development of skills among workers. Work standardization also tests the ingenuity of industrial engineers and of production or operations managers at all levels. Sometimes, ideas leading to work standardization are derived from employee suggestions. More often, however, work standards become part of an adversarial process in which they, in effect, establish work rules and are thus involved in collective bargaining and other aspects of labor relations, just as is work measurement in general.

Work standardization often leads to simplification in that it presents opportunities for eliminating and consolidating unit operations of various kinds. Searching for a common approach also requires a careful definition of what needs to be done, which is also often a source of improvements.

Work standardization often resolves larger operations into smaller elements, and the eventual standard is a combination of such smaller ones. This procedure must not be used without caution, however. The issue of independence of elements is always present; the whole may or may not equal the sum of the parts. Subsequent work measurement, especially when done by methods of statistical analysis that focus on finding outliers, that is, truly abnormal results, must then be used to revise or reinterpret work standards.

One of the best known of the more formal techniques of work standardization is group technology. This is the careful description of a heterogeneous lot of machine or other piece parts with a view to discovering as many common features in materials and dimensions as can be identified. It is then possible to start a rather large lot of a basic part through the production process, doing the common operations on all of them. Any changes or additional operations required to produce the final different parts can then be made at a later stage. The economy is realized in being able to do the identical jobs at one time.

It may prove economical to do some of the common operations on all the parts even when they are not really required, rather than set the jobs up separately. Parts like shafts, cover plates, tooling components, or mounts for electronic circuitry are among the many applications of group technology. To implement this, a special form of "cellular" layout is required which parallels the branching characteristics of the production system. Opportunities also exist for this approach in the chemical processing industries in that basic mixes can be produced, with later additives or other unit operations providing the required product differentiation.

There has been considerable progress in computerized systems to facilitate group technology. The techniques are closely linked to computer-aided design and to formalized codes that permit the detailed description of many operations. This is a necessary prerequisite to later specification of the identities. In several of these systems, the total information is finally recorded on punched cards or other memory media, with some including a microfilm of the drawing of the part required.

Simpler and less formal forms of work standardization can occur just about everywhere, in both the private and public sector. Whenever a good idea is being put forth dealing with the more efficient performance of a detailed operation, it is worthwhile to study its possible application to a wider area. Conversely, if standard practice now exists in many parts of the business, it may be advantageous to apply it to activities hitherto left unstandardized.

However, the economics of work standardization must be carefully balanced against added costs such as new equipment or tooling, carrying greater inventories, and extra handling. The establishment of larger production batches does not necessarily produce economy of scale. Some forms of group technology, for instance, lose some of their advantages when a numerically controlled machine can quickly make changes in the parts it processes as it goes along.

There is also a possibility that work standardization may lead to repetitious and monotonous operations where previously there was a degree of flexibility and variety permitted; in such a case, work standardization becomes undesirable for a variety of operational and social reasons. Great caution must also be taken to see that administrative costs do not come to exceed any operational savings realized. *See* METHODS ENGINEERING; PERFORMANCE RATING.

John E. Ullman

World Wide Web

A part of the Internet that contains linked text, image, sound, and video documents. Before the World Wide Web (WWW), information retrieval on the Internet was text-based and required that users know basic UNIX commands. The World Wide Web has gained popularity largely because of its ease of use (point-and-click graphical interface) and multimedia capabilities, as well as its convenient access to other types of Internet services (such as e-mail, Telnet, and Usenet). *See* INTERNET.

Improvements in networking technology, the falling cost of computer hardware and networking equipment, and increased bandwidth have helped the Web to contain richer content. The Web is the fastest medium for transferring information and has universal reach (crossing geographical and time boundaries). It is also easy to access information from millions of Web sites using search engines (systems that collect and index Web pages, and store searchable lists of these pages). The Web's unified networking protocols make its use seamless, transparent, and portable. As the Web has evolved, it has incorporated complementary new technologies for developing online commerce, as well as video on demand and mobile Web (Web-enabled phone), to name a few.

Development. In 1989, Tim Berners-Lee and colleagues at the European Organization for Nuclear Research (CERN) created the Hypertext Transfer Protocol (HTTP), a communication protocol for transmitting linked documents between computers. HTTP is the basis for the World Wide Web and follows the TCP/IP for the client-server model of computing. In 1992, Berners-Lee developed a software program (client), called a Web browser, which could retrieve and display text documents under the protocol. The following year, Marc Andreessen at the National Center for Supercomputing Applications created a graphical Web browser called Mosaic. Mosaic allowed users to "browse" the Web using the familiar point-and-click functionality of the Macintosh and Windows operating systems. The first commercial Web browser was released in April 1994. *See* CLIENT-SERVER SYSTEM; DATA COMMUNICATIONS.

Linking documents. A hypertext, in contrast to a conventional text, does not have to be read sequentially. A hypertext contains linked documents, and the reader chooses a path or order to follow. A link or hyperlink is a highlighted (underlined, colored, and so forth) word, phrase, or image which, when chosen, sends the user to another document (text, image, audio, video, or animation). Hyperlinks are similar to references in a scientific paper or cross references in a dictionary. In electronic documents, these cross references can be followed by a mouse click, and the documents can be anywhere in the world. A document can contain various links to other documents on different computers anywhere on the Internet, permitting easy Web browsing (**Fig. 1**).

Individual documents are called Web pages, and a collection of related documents is called a Web site. All Web documents are assigned a unique Internet address called a Uniform Resource Locator (URL) by which they can be accessed by all Web browsers. A URL (such as http://www.hq.nasa.gov/office/procurement/index.html) identifies the communication protocol used by the site (http), its location [domain name or server (www.hq.nasa.gov)], the path to the server (office/procurement), and the type of document (html).

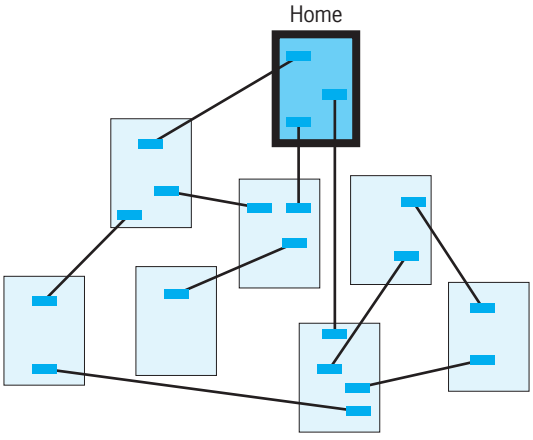


Fig. 1. In the World Wide Web, browsing pathways extend to electronic documents containing hyperlinks.

HTML. The language used to create and link documents is called Hypertext Markup Language (HTML). Markup is the process of adding information to a document that is not part of the content but identifies the structure or elements. Markup languages are not new. HTML is based on the Standard Generalized Markup Language (SGML), which became an ISO standard in 1986 and has been used in the publishing business and for technical documentation. Whereas SGML focuses on encoding the logical structure and content of a document, rather than its display formatting or even the medium in which the document will be displayed, HTML specifically encodes the format of the hypertext document for publication on the World Wide Web.

Dynamic Web pages. Though the initial format for creating a Web site was pure HTML, new and extended HTML has the ability to include programming language scripts such as common gateway interface (CGI), active server page (ASP), and Java server page (JSP), which can be used to create dynamic and interactive Web pages as opposed to just static HTML text. Dynamic Web pages allow users to create forms for transactions and data collection; perform searches on a database or on a particular Web site; create counters and track the domain names of visitors; customize Web pages to meet individual user preferences; create Web pages on the fly; and create interactive Web sites.

Extensible Markup Language (XML). XML, developed by the World Wide Web Consortium, is another derivative of SGML and is rapidly becoming the standard information protocol for all commercial software such as office tools, messaging, and distributed databases. XML is a flexible way to create common information formats and share both the format and the data on the World Wide Web, intranets, and other Web-based services. For example, automobile manufacturers might agree on a standard or common way to describe the information about a car (engine capacity, horsepower, tire specifications, mileage, transmission details, and so forth) and then describe the car information format with XML. Such a standard way of describing data would enable business users to access the manufacturers' Web pages, gather data, and then use the data for their own purpose. Any individual or group of individuals or companies that want to share information in a consistent standard way can use XML. *See* DATABASE MANAGEMENT SYSTEM.

XML is a text-generation language compared to HTML. HTML could capture only the format and the content of the text, whereas XML can capture the data related to the text. XML, like HTML, contains markup symbols (called tags) to describe the content of a Web page. For example, the XML tag <EngineCap> indicates that the data to follow is the Engine Capacity. This means that an XML file can be processed purely as data by a program, it can be stored with similar data on another computer or, just as an HTML file, it can be displayed. This is of immense use to businesses as they can now use XML for sharing data using a Web interface.

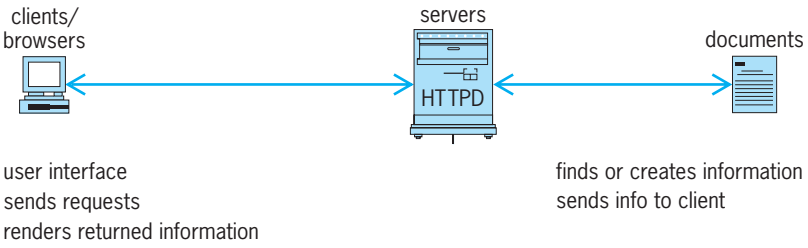


Fig. 2. Simplified view of how the Web works.

World Wide Web Consortium (W3C). The W3C is an international consortium (companies, research foundations, scientists, technologists, and universities) run by the Institut National de Recherche en Informatique et en Automatique, the Massachusetts Institute of Technology, and Keio University. The organization's purpose is to develop open (freely distributed) standards so that the Web can evolve in a single direction instead of among competing standards. The W3C is the chief standards body for HTTP and HTML. It develops interoperable technologies (specifications, guidelines, software, and tools) to lead the Web to its full potential as a forum for information, commerce, and communication.

Web architecture. In the Internet client-server system, the client (computer or program) requests documents, and the server (computer or software package) sends the requested information to the client. When a user clicks on a hyperlink, the browser (client) orders a document from the server, which delivers the requested document (or an error message if it is unavailable). The browser then presents the document to the user (**Fig. 2**). Browsers exist for most popular operating systems (such as Apple Macintosh, Microsoft Windows, and Unix). *See* OPERATING SYSTEM.

The computer or program that processes document requests is called a Web server. There is another computer, called an application server (also known as an app server), that executes commands requested by the Web server to fetch data from databases. In addition, the Web server has a database (called the DB server), which is a repository of data and content. Each page of a Web site is usually stored on a server as a separate file.

Anyone can create a Web server and, for the most part, anyone can read what is online. The reason it all works is that everyone is using the same standards. However, this increases the threat of misuse of the Web by hackers or through virus attacks on Web sites. Organizations all over the world are sensitive to such harmful possibilities. Thus security has become one of the most important issues.

Security. A firewall is a set of hardware and software programs that protects the resources of a private network from users in other networks. The firewall has the ability to restrict access to the Web server by identifying the request source's Internet address. Thus it screens unauthorized access and prevents hacking, that is, inhibits outsiders from accessing private data resources. Cryptography is beginning to play a major role in Web security. Public Key Infrastructure (PKI), 128-bit encryption, and digital certificates have been developed to protect online resources. *See* COMPUTER SECURITY.

Alladi Venkatesh

Bibliography. M. Abrams, *World Wide Web: Beyond the Basics*, Prentice Hall, Englewood Cliffs, NJ, 1998; J. Nielson, *Designing Web Usability*, New Riders, Indianapolis, 2000; C. Rolland (ed.), *Information Systems in the WWW Environment*, Chapman and Hall, London, 1998; K. Sutherland, *Understanding the Internet: A Clear Guide to Internet Technologies*, Butterworth-Heinemann, Oxford, 2000.

Wrought iron

As defined by the American Society for Testing and Materials, "a ferrous material, aggregated from a solidifying mass of pasty particles of highly refined metallic iron, with which, without subsequent fusion, is incorporated a minutely and uniformly distributed quantity of slag."

This slag is a ferrous silicate resulting from the oxidizing reactions of refining, and it varies in amount from 1 to 3% in various types of final product. It is in purely mechanical association with the iron base metal, as contrasted with the alloying relationships of the metalloids present in steel.

History. Wrought iron had a dominant position in meeting the need for a forgeable, nonbrittle, ferrous metal, from before the time of historical record until the advent of the age of steel.

The early primitive methods were finally succeeded by the puddling furnace of H. Cort in 1784. Such wrought iron played a major role in the industrial revolution until the steelmaking inventions of H. Bessemer in 1856, and the open-hearth regenerative furnace of W. Siemens several years later.

Metallurgy. Whether the product is wrought iron or wrought steel, similar chemical reactions are involved in refining a molten charge of pig iron to a composition approaching pure iron. The associated metalloids are oxidized—carbon to form carbon monoxide; and silicon, manganese, and phosphorus with appropriate additions to form a liquid insoluble slag.

The vital demarcation between the puddling and steelmaking furnace was the limited temperature of the former. The puddling furnace was below the 2700°F (1482°C) needed to keep the refined iron molten. The freezing point of iron rises as refining proceeds. Practically all the 3–4% carbon in the original charge is removed, and the base metal progressively solidifies. Manual manipulation accompanied by internal reactions results in a spongy mass of virtually pure iron, impregnated with iron silicate slag. Subsequent squeezing and rolling operations eject most of the still-fluid slag; but the small percentage finally retained is elongated into the distributed threadlike structure characteristic of quality wrought iron.

In steelmaking operations, on the contrary, the high temperature maintains fluidity of both metal and

slag, with liquation and separation of the latter. The slag-free metal is poured into molds, ready for the subsequent rolling operations.

A distinguishing characteristic of wrought iron is a fragmented or irregular fracture, as contrasted with a fibrous or crystalline type in steel. Metallographic analysis shows that this results from the fiberlike slag inclusions that are finely and uniformly intermingled with the iron base metal.

The development of steelmaking processes prompted prediction of the doom of wrought iron. Puddling required hard manual labor, with low output both in unit mass and total tonnage, and accompanying high cost. However, although wrought iron once held competitive merit in certain uses, notably where corrosion- and shock-resistance were important, very little wrought iron is produced today.

Production. Efforts to improve production by replacing hand puddling with mechanical puddling were unsuccessful because of the resulting nonuniformity in different batches due to lack of control. However, the Aston process, which approached the problem on the basis of metallurgical principles, has been used successfully to make wrought iron synthetically. The process is conducted in three main steps.

1. Pig iron is refined by standard steelmaking operations (the Bessemer converter is used).

2. An iron silicate slag of desired composition is melted in a rotary furnace (iron ore and sand are the basic ingredients).

3. The metal is poured in a steady stream into a vessel containing molten slag at a temperature below the freezing point of the iron, causing a sponge ball of wrought iron to solidify.

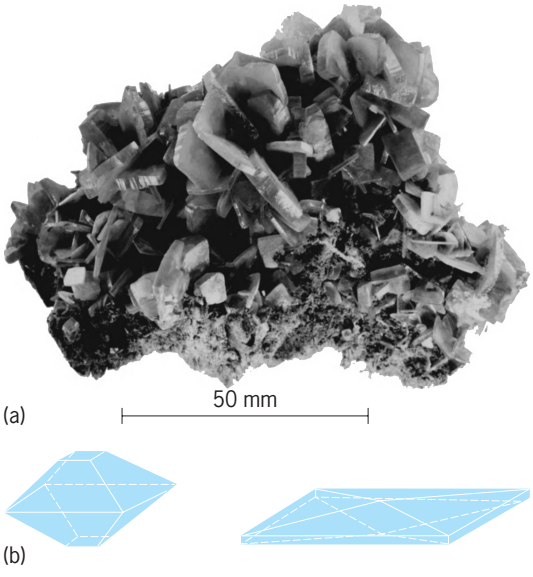
During instantaneous and progressive solidification of the metal, its dissolved gases are liberated with sufficient force to disintegrate the metal. The large volume of liquid slag is decanted and reused, and the sponge ball with its entrained slag is compacted into a large ingot while at a welding heat.

The result is wrought iron of controlled quality. Unit masses and all mechanical equipment closely parallel those of steelmaking. Diversification of product is attained, which was not feasible with the limitations of the puddling process. *See* CAST IRON; IRON ALLOYS; IRON METALLURGY; PYROMETALLURGY; STEEL; STEEL MANUFACTURE.

James Aston; John F. Wallace

Wulfenite

A mineral consisting of lead molybdate, $PbMoO_4$. Wulfenite occurs commonly in yellow, orange, red, and grayish-white crystals. They may be tetragonal,



Wulfenite. (*a*) Thin tabular crystals on limonite, Organ Mountains, New Mexico (*American Museum of Natural History specimen*). (*b*) Crystal habits (*after C. Klein and C. S. Hurlbut, Jr., Manual of Mineralogy, 21st ed., John Wiley and Sons, 1993*).

tabular (see **illus.**), or pyramidal, with a luster from adamantine to resinous. Wulfenite may also be massive or granular. Its fracture is uneven. Its hardness is 2.7–3 and its specific gravity 6.5–7. Its streak is white. It is easily fusible and is decomposed by hydrochloric or nitric acid with the separation of molybdic oxide.

Wulfenite occurs as a secondary mineral in the oxidized zone (of veins) of lead deposits associated with pyromorphite, cerussite, vanadinite, and other oxide zone minerals such as goethite and calcite.

Wulfenite is found in numerous localities in the western and southwestern United States. Brilliant orange tabular wulfenite crystals up to 2 in. (5 cm) in size have been found from the Red Cloud and Hamburg mines in Yuma County, Arizona. *See* MOLYBDENUM.

Edward C. T. Chao

Wurtzilite

A black, infusible carbonaceous substance occurring in Uinta County, Utah. It is insoluble in carbon disulfide, has a density of about 1.05, and consists of 79–80% carbon, 10.5–12.5% hydrogen, 4–6% sulfur, 1.8–2.2% nitrogen, and some oxygen.

Wurtzilite is derived from shale beds deposited near the close of Eocene (Green River) time. The material was introduced into the calcareous shale beds as a fluid after which it polymerized to form nodules or veins. *See* ASPHALT AND ASPHALTITE; IMPSONITE.

Irving A. Breger

X-ray astronomy

The study of x-ray emission from extrasolar sources, including virtually all types of astronomical objects from stars to galaxies and quasars. This space science requires balloons, rockets, or space satellites to carry experiments above most (or all) of the Earth's atmosphere, which would otherwise absorb the radiation. The x-ray region of the electromagnetic spectrum extends from wavelengths of about 10 picometers to a few tens of nanometers, with shorter wavelengths corresponding to higher-energy photons (1 nm corresponds to about 1000 eV). X-ray astronomy is traditionally divided into broad bands—soft and hard—depending on the energy of the radiation being studied. Observations in the soft band (below about 10 keV) must be carried out above the atmosphere, while hard x-ray observations can be made at high altitudes achievable by balloons. There is a limit to the observable spectrum at low energies due to absorption by neutral interstellar hydrogen gas. This low-energy cutoff depends on viewing direction, but can be below 250 eV. In the directions toward the galactic center, the cutoff can be above 2000 eV. At high energies (short wavelengths), space remains transparent through the entire x-ray spectrum. *See* ELECTROMAGNETIC RADIATION; ROCKET ASTRONOMY; X-RAYS.

X-ray observatories. In 1962, instruments aboard a rocket flight detected the first nonsolar x-ray source, later identified with a star in the constellation Scorpius (Sco X-1). The first exploratory phase in x-ray astronomy was carried out with sounding rockets and balloon flights during the late 1960s. During the early 1970s *SAS-A* (or *Uhuru*) and other satellites conducted the first full-sky surveys for x-ray emissions.

Einstein Observatory. In 1978 a major advance came to the field with the launch of the *Einstein Observatory (HEAO 2)*. This satellite introduced the use of focusing high-resolution optics to x-ray astronomy. The ensuing increase in sensitivity and the new ability to obtain images resulted in a qualitative change in the scope of the field. The images obtained with the *Einstein Observatory* during its $2^1/_2$ year operational lifetime had a major impact on many areas of study. *See* X-RAY TELESCOPE.

ROSAT. The *Roentgen Satellite (ROSAT)*, which was launched in 1990 and operated for almost 9 years, surveyed the entire sky with higher sensitivity than had been previously achieved; one result was an all-sky catalog of sources with nearly 125,000 objects, a more than twenty-fold increase in the number of known x-ray sources. *ROSAT* was also used to observe specific sources or regions of the sky for searches of extremely faint objects. The observatory used two imaging detectors. One was a high-resolution device similar to the *Einstein Observatory* High Resolution Imager (HRI), which could be moved into the focus of the x-ray telescope to obtain pictures with the best sharpness. A second detector, the Position Sensitive Proportional Counter (PSPC), provided a larger-sized image with less resolution but more information on the energy content of the x-rays from a source.

ASCA. Japan's *ASCA* (*Advanced Satellite for Cosmology and Astrophysics*, formerly *ASTRO-D*) spacecraft was launched into orbit in 1993 and operated for $7^1/_2$ years. It used lightweight modest-resolution x-ray optics that functioned at higher energies than the *Einstein* or *ROSAT* telescopes, and had advanced-technology charge-coupled-device (CCD) detectors that provided both good image quality and improved spectral resolution over previous imaging detectors. These detectors provided detailed information on the energy content of many types of x-ray sources, allowing the physical conditions at these objects to be measured. *See* CHARGE-COUPLED DEVICES.

RXTE. In late 1995 the National Aeronautics and Space Administration (NASA) launched the *Rossi X-Ray Timing Explorer (RXTE)*. This satellite measures the time variations in strong x-ray sources in order to understand the fundamental sources of

energy for these objects. Sources can be observed up to fairly high energies, which is important in resolving the different physical mechanisms responsible for the observed emission.

BeppoSAX. The Italian-Dutch *BeppoSAX* satellite, launched in 1996, can observe x-ray sources at energies ranging from 0.1 to 300 keV (although imaging is restricted to energies below 10 keV), with good energy resolution. Previously, it required the combined observations of several spacecraft to record a spectrum over this very broad energy range. In addition, two wide-field cameras can monitor the long-term variability of x-ray sources and detect transient x-ray phenomena.

The satellite is equipped with a gamma-ray burst monitor, and can quickly and accurately locate the positions of the x-ray afterglows of gamma-ray bursts that are within the field of view of the wide-field cameras. This capability makes it possible to spot the optical afterglows of some of the bursts as well, and thereby locate them even more precisely. In at least one case, observation of the optical spectrum has allowed a direct determination of the distance to the burst (about 3×10^9 light-years), strengthening the case that the bursts are generally at cosmological distances. *See* GAMMA-RAY ASTRONOMY.

Chandra. NASA's *Chandra X-ray Observatory* (*CXO*), launched in July 1999, is equipped with a 10-meter-focal-length (33-foot) telescope consisting of four pairs of mirrors arranged in concentric rings. It can focus x-rays more than 10 times as sharply as any previous x-ray telescope, with an angular resolution better than one half arcsecond, comparable to that of the best ground-based optical telescopes. The combination of high resolution with a large collecting area, sensitivity to x-rays from 0.1 to 10.0 keV, and long observation times allows *Chandra* to observe extremely faint sources.

At the telescope focus, images with the best sharpness are obtained with the High Resolution Camera (HRC), which, like similar devices on *Einstein* and *ROSAT*, is based on microchannel plates. Higher energy resolution is obtained with a second detector, the Advanced CCD Imaging Spectrometer (ACIS), with 10 charge-coupled devices similar to those used on ASCA. Transmission gratings can be swung into the path of the x-rays to achieve even better energy resolution, surpassing previous instruments. The gratings consist of meshes of fine gold wires whose spacing is precisely measured, and they deflect (diffract) the x-rays though an angle that depends on the x-ray energy. High-resolution x-ray spectroscopy enables the temperature, ionization, and chemical composition of x-ray sources to be explored. *See* DIFFRACTION GRATING.

Chandra follows an elongated orbit that takes it a third of the way to the Moon every 64 hours. This enables it to spend most of its time above the radiation belts surrounding the Earth (which can give false readings and even damage instruments) and to carry out uninterrupted long-duration observations.

XMM-Newton. In December 1999, the European Space Agency launched the *X-ray Multimirror Mission*, renamed *XMM-Newton*. The spacecraft, nearly 11 m (36 ft) long, has three x-ray telescopes, each of which has 58 concentric mirrors nested just a few millimeters apart, giving them each an effective area for collecting radiation even larger than that of *Chandra*. A charge-coupled device camera, the European Photon Imaging Camera, is located at the prime focus of each of the telescopes. For high-resolution spectroscopy, two of the telescopes are each equipped with a Reflection Grating Spectrometer. The reflection grating is a mirror with narrowly spaced grooves that reflects about half of the x-rays from the telescope to a secondary focus (with its own charge-coupled device camera), but at an angle that depends on the x-ray energy. The observatory is particularly sensitive to harder x-rays. *XMM-Newton* also has an Optical Monitor, a 30-cm (12-in.) optical-ultraviolet telescope that is aligned with the spacecraft's main telescopes and is equipped with a powerful image intensifier and a charge-coupled device camera. It allows x-ray sources to be observed simultaneously at visible and ultraviolet wavelengths. *XMM-Newton* follows a highly elliptical orbit, similar to that of *Chandra*.

Types of sources. X-ray astronomy is traditionally divided into specialties depending upon the origin of sources—galactic or extragalactic—and the energy range of observations—soft (below 10 keV) and hard (above 10 keV). Most observations have been in the soft band. Sky surveys, particularly the *ROSAT* all-sky survey, have located tens of thousands of sources at a sensitivity of about 1/100,000 of the strength of the brightest source, Sco X-1. This is comparable to a survey of the sky in visible light which extends from the brightest stars to about 12th-magnitude stars. There are over 1,000,000 stars in this range of brightness compared with the 125,000 x-ray sources detected. Some of these sources are concentrated along the galactic equator. Such a concentration corresponds to objects in the Milky Way Galaxy, particularly in the disk, which contains most of the galactic stars and the spiral arms. Other x-ray sources are uniformly spread over the sky, associated mainly with extragalactic objects, such as individual galaxies, clusters of galaxies, and quasars. *See* MILKY WAY GALAXY.

Galactic sources. Galactic x-ray sources have been identified with different types of unusual objects. With the advent of the *Einstein Observatory* and the great increase in sensitivity obtained by using focusing x-ray optics, it became possible to detect x-ray emission from stars similar to the Sun.

Normal stars. X-ray emission from the hot corona of the Sun was detected as early as the 1950s by using sounding rockets. The level of x-ray emission from the Sun is about a million times less than the emission in visible light, with occasional increases associated with solar flares. Low levels of x-ray emission have been detected from all types of normal stars. This surprising result has revived interest in models of stellar structure and evolution, which must now be reevaluated. In these sources the x-ray emission is most likely coming from the corona of the stars, with energy being transported from the stellar interior, through

the surface, and into the extended atmosphere. In all cases, the x-ray emission is a small fraction of the total luminosity of the star and is mainly interesting with regard to the problem of how energy is transported from the stellar interior to the surface. *See* SOLAR CORONA; STAR; SUN.

Supernova remnants. The Crab Nebula, which was the first nonsolar x-ray source to be identified with a specific celestial object, is a supernova remnant left over from the explosive death of a star. This particular supernova remnant is about 900 years old and contains a rapidly rotating neutron star at its center as well as a nebula consisting of hot gas and energetic particles. The neutron star is the remaining core of the star which exploded, and its rotation causes energetic electrons trapped in the magnetic fields of the nebula to radiate pulses of electromagnetic energy in radio waves, visible light, and even x-rays. About 10% of the x-ray emission from the Crab is pulsed radiation from the neutron star; the rest is extended emission associated with the nebula. *See* CRAB NEBULA.

Many other galactic supernova remnants have been detected as x-ray sources, and there have also been detections of supernova remnants in the nearest neighboring galaxies, especially in the Large Magellanic Cloud. The youngest known supernova remnant in the Milky Way Galaxy, Cas-A in the constellation Cassiopeia, is about 350 years old. The expanding shell of the explosion can easily be seen in x-ray images (see **illus.**). Analysis of the spectrum of x-ray emission shows that heavy elements, formed by nucleosynthesis within the precursor star, have been blown off during the explosion and are mixing into the interstellar medium. Observations by the *Chandra X-ray Observatory* revealed for the first time a point source near the center of Cas A which is believed to be the neutron star produced in the supernova event. Chandra observations show that the concentrations of heavy elements are arranged in a manner that echoes the ordering of the elements in the shell-burning precursor star. *See* NEUTRON STAR; NUCLEOSYNTHESIS; PULSAR; SUPERNOVA.

X-ray stars. While normal stars radiate some energy in the x-ray band, most of their luminosity is output in the visible region of the spectrum. X-ray surveys have discovered a class of objects where the majority of energy is radiated in the x-ray portion of the electromagnetic spectrum. These are known as x-ray stars. The majority of galactic x-ray stars are located in the plane of the Milky Way Galaxy, mostly within the spiral arms. This location coincides with the distribution of bright young stars in the Galaxy, and it is very likely that this class of stars goes through a stage of copious x-ray emission as the stars evolve.

Some x-ray stars have been located in the spherical halo of the Galaxy, some far from the plane, and in globular clusters. Both regions are known to contain the oldest stars in the Galaxy, and the x-ray sources there may be representative of a different class than those in the plane. Very accurate x-ray locations, followed by optical identifications and detailed studies, may help to separate x-ray stars into classes and identify different phenomena responsible for their emission. There are also transient sources which suddenly appear and reach high levels of emission rapidly, then slowly fade over periods of weeks to months. This is similar to the behavior of optical novae. *See* NOVA.

***Chandra X-ray Observatory* High-Resolution Camera image of the supernova remnant Cas A. The bright point near the center of the remnant is believed to be the neutron star that is produced during the core collapse of the star that exploded. (*S. S. Murray*)**

Compact sources and binary systems. With accurate locations, it has been possible to identify x-ray stars with binary star systems. The combination of x-ray and optical data leads to the conclusion that these systems usually consist of a relatively normal visible star and one subluminous star in a gravitationally bound orbit about each other. The x-ray star is often a collapsed object, such as a white dwarf, neutron star, or even a black hole. Such compact sources form the majority of galactic emitters.

These sources are further classified by their variability, ranging from time scales as long as years to as short as milliseconds. Some sources have periodic variations resulting from the orientation of their orbits in binary star systems relative to the Earth. When an x-ray-emitting star orbits around a normal star, and is periodically blocked from view, eclipses in the x-ray emission are observed. In some instances the x-rays appear pulsed, with a period of a few seconds or less. This is similar to observations of radio pulsars and is believed to be due to the rotation of a neutron star beaming x-rays toward the Earth.

These eclipsing, x-ray pulsars provide excellent natural laboratories for studying the effects of extremely high gravitational and magnetic fields. The theoretical model for x-ray emission in these systems consists of matter being transferred from the normal star to the compact object. This process, known as accretion, usually leads to the creation of a disk of infalling material spiraling down toward the surface of the compact star. During infall the matter reaches very high temperatures as gravitational energy is converted to heat, which is then radiated as x-rays. The

details of matter transfer, heating, and radiation are active areas of astrophysical research, involving complex relationships between the very strong gravity and magnetic fields that are present in these systems. It is believed that similar processes are involved regardless of whether the compact object is a white dwarf, neutron star, or black hole. However, different types of detailed behavior are expected for each of these objects. Rotating neutron stars lead to x-ray pulsars. For white dwarfs, there appear to be fluctuations in x-ray intensity on time scales of hours to days that might be indicative of changes at the surface of the star where accreted material is collecting, and then flashing in a burst of thermonuclear energy release. In the case of black hole candidates, there are variations in x-ray intensity at extremely short time scales that indicate very small regions of x-ray emission.

Detailed studies of the orbital motions of binary star systems, using both optical and x-ray information, allow estimates to be made of the masses of the total binary system and its component stars. In this manner, it is possible to place a range on the mass of the system's unseen compact object. Theoretical limits on the maximum allowable mass for a white dwarf and neutron star can then be used to help classify the compact object. In this way, for example, a black hole has been identified as the most likely candidate for the x-ray source Cyg X-1, in the constellation Cygnus.

Another type of time behavior is represented by the so-called bursters. These sources emit at some constant level, with occasional short bursts or flares of increased brightness. The bursts last for several minutes, and often there are long quiescent periods followed by active intervals. These outbursts may well be instabilities in the x-ray emission processes associated with this type of source. *See* BINARY STAR; BLACK HOLE; CATACLYSMIC VARIABLE; STELLAR EVOLUTION; WHITE DWARF STAR.

Extragalactic sources. As with galactic x-ray astronomy, virtually all types of objects have been detected as x-ray sources as a result of the increased sensitivity available. These distant objects are radiating enormous quantities of energy in the x-ray band, in some cases more than visible light. Among the more interesting types of objects detected have been apparent normal galaxies, galaxies with active nuclei, radio galaxies, clusters of galaxies, and quasars.

Galaxies and quasars. The nearest spiral galaxy is M31, the Andromeda Galaxy. Imaging x-ray telescopes resolve it into individual stellar sources. The distribution and association with galactic features is similar to the Milky Way. There appears to be very faint emission associated with M31 as a whole, probably due to hot interstellar gas, again similar to diffuse galactic emission detected in the Milky Way. There is also a weak x-ray source located at the center of M31 where a supermassive black hole with about 3×10^7 times the mass of the Sun has been detected. The low x-ray brightness of the nucleus of M31 is an example of a weak nuclear galactic source. *See* ANDROMEDA GALAXY.

Observations of seemingly normal galaxies have led to the discovery of a new class of x-ray source which might be labeled optically dull. These objects emit unusually large fluxes of x-rays, much more than the summed emission of their stars, but appear to be otherwise normal in their optical and radio properties. The mechanisms that produce these x-rays are not well understood. In some cases, detailed studies of optical emission from such galaxies have led to the discovery of signs of nuclear activity in the form of weak emission lines.

In active nuclei galaxies (those with strong optical emission lines and nonthermal continuum spectra), x-ray emission is usually orders of magnitude in excess of that from normal galaxies. It is clearly associated with the galaxy itself and not the summed emission from stars. In most cases the emission comes from the galactic nucleus, and must be confined to a relatively small region on the basis of variability and lack of structure at the current observational limits of a few seconds of arc. Many astrophysicists believe that galactic nuclei are the sites of massive black holes, at least 10^6–10^9 times the mass of the Sun, and that the radiation from these objects is due to gravitational energy released by infalling material. Some scientists speculate that this is a common feature of all galaxies (perhaps with the massive black holes being as small as 10^6 solar masses in so-called normal galaxies), and that the broad range of properties, such as x-ray luminosity, reflects the size of the black hole and availability of infalling matter. Thus, normal galaxies have dormant nuclei where there is little or no material available, while Seyfert galaxies are quite active and quasars represent the extreme case of this mechanism.

In a unified picture of active galaxies, many of the observable properties depend on the viewing angle of the observer, as well as the size of the nuclear black hole. In this model, the black hole lies at the center of an accretion disk (shaped like a doughnut), and beyond this component other clouds of material that give rise to the optical emission lines. If the observer's line of sight is along the axis of the black hole and accretion disk, then the intense x-ray source is seen directly, whereas if the viewing angle is in the plane of the accretion disk, the intense x-ray source is not seen and the object will appear to be much weaker. X-ray observations are particularly relevant in understanding the nature of galactic nuclei, since the high-energy emission is a more direct probe of the basic energy source than optical, infrared, or radio. *See* QUASAR.

Clusters of galaxies. Clusters of galaxies are collections of hundreds to thousands of individual galaxies which form a gravitationally bound system. They are among the largest aggregates of matter in the universe (along with clusters of clusters), and can be detected at very large distances. This allows them to be studied at early epochs in the development of the universe. The space between galaxies in such clusters has been found to contain hot (approximately 10^7–10^8 K) tenuous gas which glows in x-rays. The mass of matter in this gas is equal to (or exceeds)

the mass of the visible galaxies; it contains heavy elements which were not present in the primordial mixture from which stars and galaxies are believed to have formed.

Studies of clusters can be used to help determine the order of formation in the early universe. There are two competing scenarios for the relationship between the time that objects form and their sizes. In one, the largest structures form first and then these evolve to form clusters of galaxies and galaxies. In the other, galaxies form first and then clump together to form the clusters and superclusters. X-ray observations of clusters of galaxies that have subclustering are taken to be examples of cluster formation in process, giving support to the hierarchy of small-to-large formation processes. However, optical observations of the distribution of galaxies over large regions of the universe show large-scale structures that resemble bubbles and voids, with the galaxies tending to be found on the surfaces of these bubbles. These observations are taken to be evidence for the initial formation of large-scale structures followed by smaller objects. Theoretical models tend to include aspects of both scenarios, but further observational data will be needed to clarify the ideas and test the models in more detail. *See* COSMOLOGY; GALAXY, EXTERNAL; UNIVERSE.

X-ray background. In addition to observing a bright source later identified as Sco X-1, the 1962 rocket experiment discovered a uniform background of x-ray emission around the sky. This observation was confirmed by subsequent experiments, and the background was found to contain at least two components. At low energies the background is dominated by galactic emission, while above about 1 keV the radiation comes from beyond the Milky Way Galaxy. The galactic background appears to be diffuse and varies depending upon the viewing direction. It is most likely due to hot (10^6 K) interstellar gas that was heated by a supernova explosion. Because the Milky Way is not very transparent at low x-ray energies, this hot gas is probably nearby (within a few hundred light-years).

The harder component has two possible sources. Either there is a truly diffuse source of emission, such as hot gas or very high energy particles, or the background is the superposition of many discrete sources associated with various extragalactic objects. The spectrum of the background has a smooth shape that can be well approximated by the combined spectra of several types of extragalactic sources or by that expected from a very hot gas. Energy considerations rule out high-energy particles as a source of the background, but large numbers of extragalactic objects have been found by using source counts. Thus, observed classes of individual sources account for at least 80%, and perhaps all, of the background observed. *Chandra* and *XMM-Newton* observations have confirmed this accounting. *ROSAT* and *Chandra* together resolved 90% of the background at energies up to 2 keV into distinct sources, and *Chandra* and *XMM-Newton* resolved most of the background between 2 and 10 keV. All these results limit any remaining contribution from a diffuse hot gas to be, at most, a few percent, and this limits the total mass of hot gas that can be present in the universe to an amount less than that required for gravitational closure. That is, there is not enough light-emitting or normal matter presently identified in the universe to overcome the initial expansion of the big bang and cause a latter contraction. If such matter exists, it must be some heretofore-unobservable form, called dark matter, such as very dim stars; small, isolated black holes; or exotic particles.

The source-count data from deep x-ray surveys indicate that the mix of different classes of extragalactic sources varies with their distance. The strongest and closest extragalactic sources consist mainly of clusters of galaxies and active galaxies such as Seyfert types. At fainter fluxes and thus greater distances, there are mainly active galaxies such as quasars. These observations provide evidence for rapid evolution of quasars in the sense that they were more common in the earlier universe than in the present epoch. *See* ASTROPHYSICS, HIGH-ENERGY.

Stephen S. Murray

Bibliography. B. Aschenbach., H. M. Hahn, and J. E. Trumper, *The Invisible Sky: Rosat and the Age of X-ray Astronomy*, 1998; X. Barcons and A. C. Fabian (eds.), *The X-ray Background*, 1992; W. Lewin, J. Van Paradijs, and E. Van Den Heuvel (eds.), *X-ray Binaries*, 1995; L. Scarsi et al. (eds.), *The Active X-ray Sky: Results from BeppoSAX and RXTE*, 1998; F. D. Seward and P. A. Charles, *Exploring the X-ray Universe*, 1995; W. Tucker and K. Tucker, *Revealing the Universe: The Making of the Chandra X-ray Observatory*, 2001.

X-ray crystallography

The study of crystal structures by x-ray diffraction techniques. The prediction in 1912 by the German physicist Max von Laue that crystals might be employed as natural diffraction gratings in the study of x-rays was experimentally verified in the same year by W. Friedrich and P. Knipping, who obtained diffraction patterns photographically by the so-called Laue method. Almost immediately after (1913), W. Lawrence Bragg not only successfully analyzed the structures of sodium chloride and potassium chloride by Laue photographs but also developed a simple treatment of x-ray scattering by a crystal (the Bragg law) which proved much easier to apply than the more complicated but equivalent Laue theory of diffraction. The availability of the first x-ray spectrometer, constructed by his father, William H. Bragg, as well as the substitution of monochromatic (single-wavelength) for polychromatic x-ray radiation, enabled W. L. Bragg to determine a number of simple crystal structures, including those of diamond; zincblende, ZnS; fluorspar, CaF_2; and pyrites, FeF_2.

For the next 50 years x-ray crystallography was used to determine only a limited number of crystal structures, with relatively low accuracy. This

situation has completely changed since the 1960s due to a combination of several factors: (1) the development of the so-called direct methods for the determination of both centrosymmetric and noncentrosymmetric crystal structures in a straightforward and rapid fashion; (2) the development of computer-controlled automatic x-ray diffractometers for rapidly collecting relatively precise intensity data (at low temperatures); and (3) the availability of highly sophisticated computers for conveniently handling all numerical calculations in the solution and least-squares refinement of each crystal structure. Advantages of modern single-crystal x-ray structural analysis are (1) it is the most powerful method available for the detailed characterization of any crystalline material (including macromolecules such as proteins and nucleic acids); (2) it provides highly accurate stereochemical information (including absolute configurations for optically active species) which in general is complete and unambiguous; and (3) it enables the structures of compounds containing 100–200 independent atoms to be determined very rapidly, that is, in 1–2 days. In the 1960s, 1–2 years were commonly required to work out the structures of relatively simple compounds with less than 20 independent atoms. The tremendous impact of modern x-ray crystallography as an analytical structural tool is apparent from the fact that thousands of structures of organic, inorganic, organometallic, and biological compounds (both small and large) are determined each year by x-ray diffraction. For the theoretical and experimental aspects of x-ray diffraction *see* X-RAY DIFFRACTION.

Bragg's law. Structurally, a crystal is a three-dimensional periodic arrangement in space of atoms, groups of atoms, or molecules. If the periodicity of this pattern extends throughout a given piece of material, one speaks of a single crystal. The exact structure of any given crystal is determined if the locations of all atoms making up the three-dimensional periodic pattern called the unit cell are known. The very close and periodic arrangement of the atoms in a crystal permits it to act as a diffraction grating for x-rays. W. Lawrence Bragg treated the phenomenon of the interference of x-rays with crystals as if the x-rays were being reflected by successive parallel equidistant planes of atoms in the crystal. His important equation relating the perpendicular spacing d of lattice planes in a crystal, the glancing angle θ_{bkl} of the reflected beam, and the x-ray wavelength λ is Eq. (1). This expression provides the basic condition that the difference in path length for waves reflected from successive planes must be an integral number of wavelengths $n\lambda$ in order for the waves reflected from a given set of lattice planes to be in phase with one another. Instead of referring to the nth order of a reflected beam, modern crystallographers customarily redefine Bragg's equation as Eq. (2), where d_{bkl} represents the perpendicular interplanar distance between adjacent lattice planes having the Miller indices (hkl). According to this viewpoint any nth-order diffraction maxima for waves reflected from a set of planes (hkl) with spacing d_{bkl} is equal to a first-order reflection due to a parallel set of planes (nh, nk, nl) with the perpendicular distance given by Eq. (3). For a discussion of Bragg's relation *see* X-RAY POWDER METHODS.

$$n\lambda = 2d \sin\theta_{bkl} \tag{1}$$

$$\lambda = 2d_{bkl} \sin\theta_{bkl} \tag{2}$$

$$d_{nb,nk,nl} = \frac{d_{bkl}}{n} \tag{3}$$

For a crystal with known unit-cell size and shape and with arbitrary orientation in a parallel beam of x-rays of known wavelength, several questions must be answered. First, it must be ascertained what plane, if any, is obeying Bragg's law and is reflecting the rays. Second, the direction of the reflected x-ray must be accurately measured. The Bragg law treatment does not permit these questions to be answered readily, and for this reason the concept of a reciprocal lattice model was introduced. *See* CRYSTAL STRUCTURE; CRYSTALLOGRAPHY.

Reciprocal lattice. The development and application of the reciprocal lattice to x-ray crystallography have been primarily the results of work by P. P. Ewald (1913 and 1921), J. D. Bernal (1926), and M. J. Buerger (1935). In general, a reciprocal lattice consists of a three-dimensional array of points which is related to the crystal lattice (commonly called the direct lattice) in that each set of planes (hkl) in the crystal lattice is represented in reciprocal space by a point denoted by the coordinates hkl (without parentheses). **Figure 1** illustrates in two dimensions the reciprocal lattice geometrically produced in the following way from a unit crystal cell. An origin is chosen and given the coordinates 000. To every set of parallel planes of Miller index (hkl) in direct space, a reciprocal lattice vector H_{hkl} which is perpendicular to this set of planes is constructed

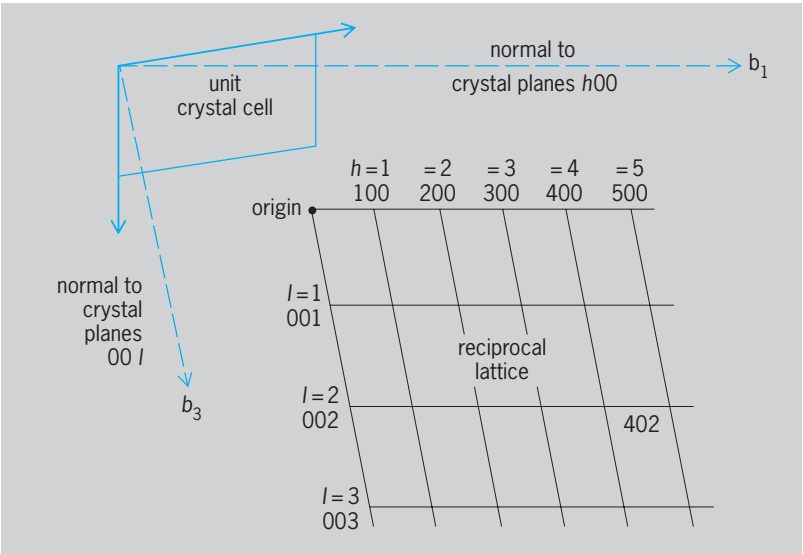


Fig. 1. Reciprocal lattice, shown in two dimensions. Each point in this lattice is reciprocal to a set of planes in the direct lattice.

from the reciprocal lattice origin at a distance inversely proportional to the interplanar spacing d_{hkl}. Thus, the set of parallel direct-lattice planes ($\bar{4}02$) with an interplanar spacing $d_{\bar{4}02}$ is represented in reciprocal space in Fig. 1 by the point $\bar{4}02$. The normal vector $H_{\bar{4}02}$, which is directed along the perpendicular of the ($\bar{4}02$) direct-lattice planes from the origin of the reciprocal lattice to the point with coordinates $\bar{4}02$, is of length $|H_{\bar{4}02}| = K/d_{\bar{4}02}$, where K is an arbitrary constant. The value of K which simply scales the reciprocal lattice is normally taken as unity or as the wavelength λ; for purposes of this discussion it is taken as 1. It can be shown that this procedure applied to a three-dimensional periodic crystal lattice will result in a three-dimensional reciprocal lattice. Each point of coordinates *hkl* in this reciprocal lattice will be reciprocal toa set of planes (*hkl*) in the direct crystal lattice.

Sphere of reflection. The geometrical interpretation of x-ray diffraction from a crystal can be best interpreted by the use of a "sphere of reflection" in reciprocal space (**Fig. 2**). A sphere of radius $1/\lambda$ (corresponding to $K = 1$) is drawn with the direction of the primary x-ray beam (that is, both the incident and transmitted rays) denoted by the unit vector s_0 assumed to travel along a diameter. The crystal C is imagined to be at the center of this sphere of reflection. The origin 000 of the reciprocal lattice is placed at the point O, where the transmitted beam emerges from the sphere of reflection. A diffracted beam will be formed if the surface of the sphere of reflection intercepts any reciprocal lattice point (other than the origin O). In Fig. 2 the reciprocal lattice point P lies on the surface of the sphere of reflection. Hence, two important results follow: (1) the set of crystal planes (*hkl*) to which this point is the reciprocal obeys Bragg's law and reflects the incident x-ray beam; (2) the direction of the diffracted rays will be from the point C at the center of the sphere to where the point P lies on the surface of the sphere. The angle of diffraction between the diffracted beam s and the primary beam s_0 is $2\theta_{hkl}$.

The reciprocal lattice, by virtue of its definition, is tied to the actual lattice insofar as orientation is concerned. As the crystal (and therefore the crystal lattice) is turned about an axis of rotation, the reciprocal lattice, in turning a similar angle about a parallel axis through its origin O, passes through the sphere of reflection (which is fixed for the primary x-ray beam that is stationary relative to the movement of the crystal). In the rotating crystal method all diffraction maxima that can possibly be recorded by any chosen x-ray wavelength λ must be represented by those points *hkl* of the reciprocal lattice that cut the sphere of reflection. These points of possible diffraction maxima all lie within a sphere of radius $2/\lambda$ called the limiting sphere. This geometrical interpretation of Bragg's law enables one to understand readily both the geometry of reflection of x-rays and the manipulation of all the single-crystal cameras and diffractometers now in use.

Crystal structure determination. A modern structural analysis by x-ray diffraction usually involves the following sequential steps: (1) selection and preliminary x-ray characterization of a suitable single crystal followed by collection of intensity data normally with a diffractometer; (2) structural determination by solution of the phase problem to give a correct trial structure; (3) least-squares refinement of the trial structure; and (4) calculation of structural parameters and presentation and display of results.

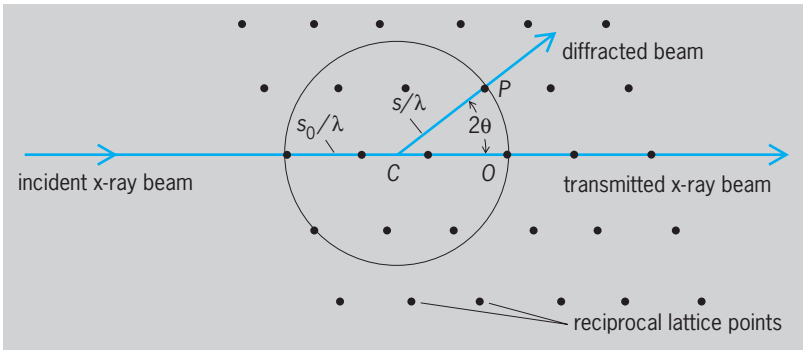


Fig. 2. Sphere of reflection in reciprocal space. Dots represent reciprocal lattice points. Diffraction occurs when any reciprocal lattice point *hkl* cuts the sphere of reflection, as at point *P*.

Collection of intensity data. Since x-rays are scattered by the electrons of the atoms, the intensity of each diffracted beam depends on the positions of the atoms in the unit cell. Any alteration in atomic coordinates would result in changes of the intensities relative to one another. Hence, the first important step in a crystal-structure determination is concerned with the collection of the intensity data by a recording technique which effectively measures the intensity at each reciprocal lattice point.

In the early days of x-ray crystallography, the classical Bragg x-ray spectrometer was used with an ionization chamber to collect data on the intensity of x-ray reflections. It was soon found possible and more convenient to use photographic recording for such data collection. Until the early 1960s the great majority of structural determinations were based on photographically recorded intensities which were usually visually estimated. Nevertheless, this process of photographic data collection was time-consuming, and normally several months were needed to record, judge, average, and scale the intensities of three-dimensional film data. The estimated level of precision of the relative intensities obtained photographically was 15–20%, which was generally sufficient for the solution of stereochemical and conformational problems but was not sufficient for a reliable determination of atomic thermal motion or bonding. Since about 1953 highly stabilized x-ray diffraction units and sensitive, reliable detectors such as scintillation counters have been developed. This modern instrumentation, coupled with the demand for making the intensity collection from crystals both more rapid and more accurate (especially that from proteins because of their instability and large number of data to be recorded), has resulted in a return to direct recording by counter techniques. The commercial availability of automatic

diffractometers which use computerized circuitry to synchronize the movement of the crystal and the detector has effectively revolutionized data collection from crystals such that upward of 1000–3000 diffracted beams can be obtained daily with errors normally less than 3%. *See* SCINTILLATION COUNTER.

Prior to a preliminary x-ray analysis, a single crystal (normally 0.05–0.5 mm in a given dimension) is optically chosen, mounted either on a glass fiber with epoxy or inside a thin-walled glass capillary, and placed on a goniometer head attached to the diffractometer. The preliminary x-ray characterization involves both optical and x-ray alignment with automatic indexing of diffracted beams and selection of a unit cell, from which lattice constants and an orientation matrix are calculated. After selection of data collection parameters, the computer controlling the diffractometer uses the orientation matrix, the *hkl* indices, and wavelength of the monochromatic x-ray beam (normally Cu$K\alpha$ or Mo$K\alpha$ radiation) to calculate angle settings for each diffracted beam to be measured. The intensity $I(hkl)$ and its estimated standard deviation $\sigma(I)$ are measured individually for each reflection with a sensitive scintillation counter detector. Data collection involves the computer-recording of the $I(hkl)$ and $\sigma(I)$ for all independent h,k,l reflections until an entire diffraction pattern has been collected. For a typical organic compound with 20 nonhydrogen atoms, the diffraction pattern may possess approximately 2000 independent reflections. Data reduction involves the correction of the raw $I(hkl)$'s for background, possible radiation-induced crystal decay, Lorentz-polarization effects, and crystal absorption effects, after which the corrected intensities are converted to observed structure factor amplitudes $|F_{hkl}|_o$. The space group symmetry of the unit cell is deduced by examination of the tabulated systematic absences of the $I(hkl)$'s.

Solution of phase problem. Although the size and shape of the unit cell determine the geometry of the diffraction maxima, the intensity of each reflection is determined by the number, character, and distribution of the atoms within the unit cell. The second stage in a structural analysis is the solution of the phase relations among the diffracted beams (that is, the phase problem) from which a correct trial structure can be obtained. The intensity of each diffracted beam is related to the square of its amplitude; that is, Eq. (4)

$$I_{hkl} = k\,|F_{hkl}|_o^2 \tag{4}$$

holds, where $|F_{hkl}|_o$ is the observed structure factor amplitude. Each diffracted beam has not only a characteristic intensity but also a characteristic phase angle α_{hkl} associated with it which expresses the degree to which the diffracted beam is in plane with the other diffracted beams.

Because the electron density is a real, positive quantity which varies continuously and periodically in a crystal, the electron scattering density $\rho(xyz)$ is derivable from the three-dimensional Fourier series as in Eq. (5), where $\rho(xyz)$ represents the electron

$$\rho(xyz) = \frac{1}{V}\sum_h\sum_k\sum_l |F_{hkl}|_o \times \cos\left[2\pi(hx + ky + lz) - \alpha_{hkl}\right] \tag{5}$$

density at any point with fractional coordinates x, y, z in the unit cell of volume V. Hence, if the characteristic amplitude $|F_{hkl}|_o$ and its phase α_{hkl} for each diffracted beam are known, the electron density can be calculated at fractional grid points in the unit cell. A properly phased three-dimensional electron-density map effectively provides peaks characteristic of direct images of atoms in the unit cell. The phase problem in x-ray crystallography arises because experimental measurements yield only the magnitudes of the structure factors $|F_{hkl}|_o$ but not the phases α_{hkl}. Hence, a structural analysis involves a search for the characteristic phases to be utilized together with the observed amplitudes in order to obtain a three-dimensional electron-density map and thereby to determine the crystal structure.

The structure factor itself is related to the scattering by the atoms in the unit cell by Eq. (6), where

$$F_{hkl} = \sum\nolimits_n f_n T_n \exp\left[2\pi i(hx_n + ky_n + lz_n)\right] \tag{6}$$

x_n, y_n, z_n are the fractional coordinates of atom n along the three crystallographic axes. The f_n's are the individual atomic scattering factors which are known for an atom at rest. At zero 2θ angle of scattering for which all the electrons of an atom scatter in phase, f_n is equal to the atomic number. The T_n's are the individual modifications which cause a decrease of the f_n's as a result of thermal motion. If an atom n is assumed to vibrate equally in all directions, its thermal behavior is designated by an isotopic temperature factor B_n (in nm^2) given by Eq. (7); a large

$$T_n = \exp\left(-B_n \sin^2\theta_{hkl}/\lambda^2\right) \tag{7}$$

B_n value (greater than 0.05 nm^2) corresponds to a large thermal displacement of an atom from its mean position and hence to a small T_n value. In general, atomic thermal motion varies with direction such that the isotropic B_n of an atom n is replaced by six atomic thermal parameters which describe an ellipsoidal electron distribution. Hence, the thermal motion of an atom is conventionally represented by a probability thermal ellipsoid (that is, an atom has a given probability, usually 50%, of being inside the thermal ellipsoid). If the positions of the atoms in the unit cell are known, the complex structure factor for each diffracted beam *hkl* can be calculated and both its magnitude and phase obtained by Eq. (8).

$$F_{hkl} = A_{hkl} + iB_{hkl} \tag{8}$$

$$A_{hkl} = \sum\nolimits_n f_n T_n \cos 2\pi(hx_n + ky_n + lz_n)$$

$$B_{hkl} = \sum\nolimits_n f_n T_n \sin 2\pi(hx_n + ky_n + lz_n)$$

$$i = \sqrt{-1}$$

Consequently, the calculated structure factor amplitude $|F_{bkl}|_c$ and the associated phase α_{bkl} can be obtained from Eq. (9).

$$|F_{bkl}|_c = \sqrt{A_{bkl}^2 + B_{bkl}^2} \qquad (9)$$

$$\tan \alpha_{bkl} = \frac{B_{bkl}}{A_{bkl}}$$

For centrosymmetric crystals which have a center of symmetry located at the corner of each unit cell, the structure factor F_{bkl} simplifies to a real rather than a complex number; that is, since for each atom at x, y, z, there must exist a centrosymmetrically related atom at $-x, -y, -z$, it follows that $B_{bkl} = 0$ and $F_{bkl} = \pm A_{bkl}$. The phases of the diffracted beams then are restricted to being either completely in phase (that is, $\alpha_{\mathrm{hkl}} = 0^\circ$ corresponding to $F_{bkl} = +A_{bkl}$) or completely out of phase (that is, $\alpha_{bkl} = 180^\circ$ corresponding to $F_{bkl} = -A_{bkl}$) with one another.

Various means of deducing the phase angle of each diffracted beam have been used. Prior to the advent of direct methods, the main approach used to overcome the phase problem was based on the discovery by A. L. Patterson in 1935 that when the experimentally known quantities $|F_{bkl}|_0^2$ instead of F_{bkl} are used as coefficients in the Fourier series, the maxima in this summation then correspond not to atomic centers but rather to a map of interatomic vectors. This vector map represents a superposition of all interatomic vectors between pairs of atoms translated to the origin of the unit cell. Since the peak height for a given vector between two atoms is approximately proportional to the product of their atomic numbers, the vectors between heavy atoms (that is, those of high atomic number) usually stand out strongly against the background of heavy-light and light-light atom vectors. Consequently, the Patterson map is especially applicable to the structure determination of compounds containing a small number of relatively heavy atoms because approximate coordinates of these heavy atoms can be obtained provided the heavy-atom vectors are correctly recognized in the Patterson map. Computers have enabled the development of a number of powerful techniques in unraveling the Patterson vector map in terms of atomic coordinates; these include multiple superpositions of parts of the Patterson map and "image-seeking" with known vectors.

Direct methods of phase determination presently constitute the universal approach to solving the phase problem for both centrosymmetric and noncentrosymmetric crystal structures. Direct methods make use of probability theory to generate an adequate set of phases by consideration solely of statistical relationships among the observed structure factor amplitudes of the diffracted beams. In order to utilize direct methods, the observed structure factor amplitudes are first converted to normalized structure factor amplitudes defined by Eq. (10), where ϵ

$$|E_{bkl}| = \left(\frac{|F_{bkl}|_o^2}{\epsilon_i \sum_n f_n^2} \right)^{1/2} \qquad (10)$$

is a small integer (normally 1) that is space-group-dependent. The distribution of the $|E_{bkl}|$'s is in principle independent of the size and content of the unit cell. The application of the so-called tangent formula, first derived by J. Karle and H. Hauptman in 1956, to reflections with large $|E_{bkl}|$'s provides sets of probable phase relationships. These E-value reflections with their most probable phases are normally then utilized to calculate Fourier E-maps (that is, the $|E_{\mathrm{bkl}}|$'s are coefficients instead of the $|F_{bkl}|_o$'s) whose peak maxima are generally sufficiently resolved that they can be interpreted in terms of an approximate structure. Direct-methods computer programs such as MULTAN have been particularly successful in routinely solving the crystal structures of a wide variety of biological molecules. Although these computer programs cannot as yet be used alone to determine the structures of macromolecules such as proteins, extensive research is being carried out to apply the methodology to the structural solution of large molecules.

Fourier electron-density and ΔF maps. Any resulting trial structure consisting of initial coordinates for some of the atoms is often sufficient for location of the other atoms by the application of the method of successive Fourier electron-density maps. The phases calculated from the initial parameters of the presumably known atoms, together with the observed $|F_{bkl}|$, are utilized to compute a density map. New coordinates are obtained not only for the peaks corresponding to these known atoms but also for the other

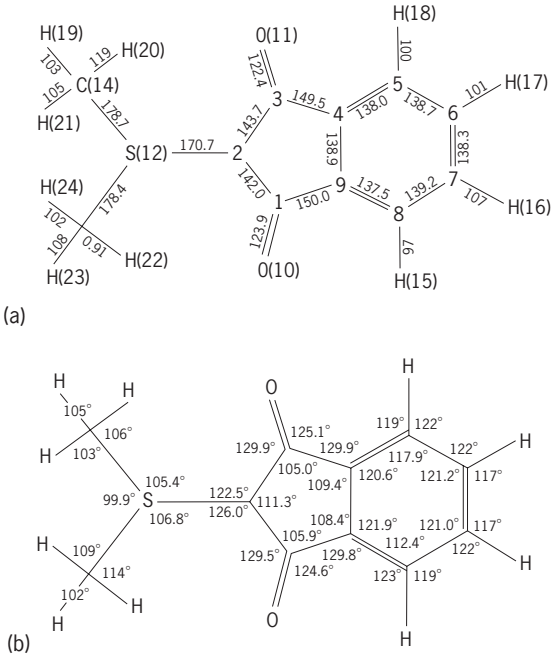


Fig. 3. Configuration of 2-dimethylsulfuranylidene-1,3-indanedione molecule which was refined by least squares to a final *R* value of 0.028 for 1170 independent diffractometry-collected *hkl* data. (*a*) Bond lengths in picometers. (*b*) Bond angles. Estimated standard deviations involving the nonhydrogen atoms are in the 0.3–0.5-pm range for distances and 0.1–0.3° range for bond angles, while those involving the hydrogen atoms are approximately 5 pm and 2.8°.

peaks which it is hoped can be interpreted from stereochemical considerations as being due to additional atoms in the structure. The Fourier process is then reiterated, with the new phases calculated from the modified coordinates of the previous set of atoms plus the coordinates of newly located atoms. If a correct distinction between the "true" and the "false" peaks is made, the electron-density function usually converges to give the entire crystal structure.

Another Fourier synthesis which is often used is a difference synthesis, $\Delta\rho(xyz) = \rho_o(xyz) - \rho_c(xyz)$, which is given by Eq. (11), where ΔF is given by Eq. (12) and α_{hkl} is the calculated phase of $|F(hkl)_c|$.

$$\Delta\rho(xyz) = \frac{1}{V}\sum_h\sum_k\sum_l \Delta F \exp(i\alpha_{hkl}) \times \exp\left[-2\pi i(hx + ly + lz)\right] \quad (11)$$

$$\Delta F = |F(hkl)|_o - |F(hkl)|_c \quad (12)$$

The particular virtue of this ΔF map is that it eliminates many of the false maxima which are associated with termination of series errors. A difference map is commonly used to locate hydrogen atoms whose positions may be obscured in a regular Fourier electron density map by the relatively heavy nonhydrogen atoms. It provides a means of subtracting out the electron density of the other atoms in the structure.

Refinement of parameters. Once the phase problem is solved and the approximately correct trial structure is known, the next step of the structural analysis involves the refinement of the positional and thermal parameters of the atoms. Normally this refinement is carried out analytically by the application of a nonlinear least-squares procedure in which a weighted quantity such as $\Sigma w[|F(hkl)|_o - |F(hkl)|_c]^2$ (where the weights w are appropriate to the experiment) is minimized with respect to the parameters. This method of refinement, which gives the best fit between the calculated and observed structure factor amplitudes, yields the most precise values for the parameters. Although there is no single reliable method for directly assessing the accuracy of a structural determination, a criterion commonly used is the unweighted reliability factor or discrepancy

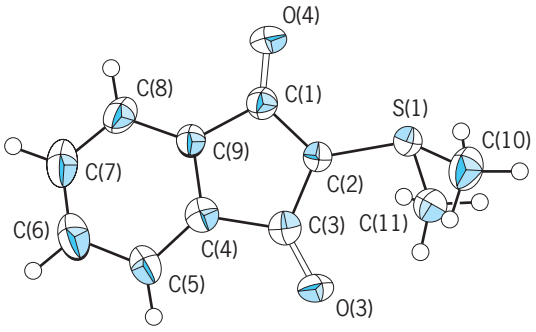


Fig. 4. Molecular configuration of the crystallographically independent $C_{11}O_2H_{10}S$ molecule. The thermal ellipsoids for the nonhydrogen atoms, which were refined with anisotropic thermal parameters, are of 50% probability.

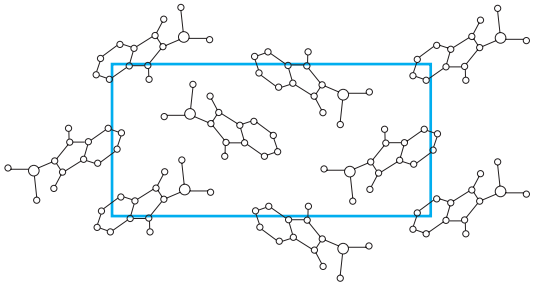


Fig. 5. Unit-cell diagram showing the arrangement of the four symmetry-related $C_{11}O_2H_{10}S$ molecules.

index R_1, defined by Eq. (13) as the summation of

$$R_1 = \frac{\sum_{hkl}\left||F(hkl)|_o - |F(hkl)|_c\right|}{\sum_{hkl}|F(hkl)|_o} \quad (13)$$

the absolute difference in the observed and calculated structure amplitudes divided by the summation of the observed amplitudes. The better the structure, including the atomic coordinates and thermal parameters, is known, the more nearly will the calculated amplitudes agree with the observed ones, and hence the lower will be the R_1 value. Discrepancy values of "finished" modern structural analyses found in the literature vary from approximately 0.12 to less than 0.02, depending upon a number of factors, such as the complexity of the structure and the number and quality of the data obtained.

Before the development of large electronic computers the determination and refinement of a crystal structure of more than about 20 atoms was not feasible. Nowadays, single-crystal analyses of uncomplicated structures may require the location of as many as 200 atoms, but there is no high correlation between the complexity of the structural determination and the number of atoms involved.

Calculations and presentations of results. The output of the least-squares refinement provides atomic parameters, namely, three x, y, z coordinates and either one isotropic or six anisotropic thermal parameters per atom, depending upon the thermal model used in the refinement process. These refined parameters and their estimated standard deviations may be used in the last step of the structural analysis to obtain the following information: (1) interatomic distances and bond angles with estimated standard deviations; (2) "best" least-squares planes of specified groups of atoms; (3) torsional angles; and (4) atomic thermal behavior. Interactive graphics programs on an appropriate computer terminal may then be utilized to generate various plots of the structure (**Figs. 3–5**).

Lawrence F. Dahl

Bibliography. T. L. Blundell and L. N. Johnson, *Protein Crystallography*, 1976; J. P. Glusker and K. N. Trueblood, *Crystal Structure Analysis: A Primer*, 2d ed., 1985; M. F. C. Ladd and R. A. Palmer, *Structure Determination by X-ray Crystallography*, 3d ed., 1993; P. Luger, *Modern X-ray Analysis on Single*

Crystals, 1980; G. H. Stout and L. H. Jensen, *X-ray Structure Determination: A Practical Guide,* 2d ed., 1989.

X-ray diffraction

The scattering of x-rays by matter with accompanying variation in intensity in different directions due to interference effects. X-ray diffraction is one of the most important tools of solid-state chemistry, since it constitutes a powerful and readily available method for determining atomic arrangements in matter. X-ray diffraction methods depend upon the fact that x-ray wavelengths of the order of 1 nanometer are readily available and that this is the order of magnitude of atomic dimensions. When an x-ray beam falls on matter, scattered x-radiation is produced by all the atoms. These scattered waves spread out spherically from all the atoms in the sample, and the interference effects of the scattered radiation from the different atoms cause the intensity of the scattered radiation to exhibit maxima and minima in various directions. *See* DIFFRACTION.

Some of the uses of x-ray diffraction are (1) differentiation between crystalline and amorphous materials; (2) determination of the structure of crystalline materials (crystal axes, size and shape of the unit cell, positions of the atoms in the unit cell); (3) determination of electron distribution within the atoms, and throughout the unit cell; (4) determination of the orientation of single crystals; (5) determination of the texture of polygrained materials; (6) identification of crystalline phases and measurement of the relative proportions; (7) measurement of limits of solid solubility, and determination of phase diagrams; (8) measurement of strain and small grain size; (9) measurement of various kinds of randomness, disorder, and imperfections in crystals; and (10) determination of radial distribution functions for amorphous solids and liquids.

For the study of crystal structure by x-ray diffraction techniques *see* X-RAY CRYSTALLOGRAPHY.

Diffraction Theory

When x-rays fall on the atoms of a substance, the scattered radiation is of two kinds: Compton modified scattering of increased wavelength which is incoherent with respect to the primary beam, and unmodified scattering coherent with the primary beam. Because of interference effects from the unmodified scattering by the different atoms of the sample, the intensity of unmodified scattering varies in different directions. A diagram of this variation in direction of intensity of unmodified scattering is called the diffraction pattern of the substance. This pattern is determined by the kinds of atoms and their arrangement in the sample; for simple structures the atomic arrangement is readily deduced from the diffraction pattern. *See* COMPTON EFFECT.

The atomic scattering factor f is defined as the ratio of the amplitude of unmodified scattering by an atom to the amplitude of scattering by a free electron, which scatters according to classical theory. In general, f is a real number which decreases with $(\sin\theta)/\lambda$, where θ is the grazing angle and λ the wavelength, from an initial value $f = Z$, where Z is the number of electrons in the atom. However, if the x-ray wavelength is close to an absorption edge of the atom, f becomes complex. For a definition of absorption edges *see* X-RAY FLUORESCENCE ANALYSIS.

If the electron density in the atom has spherical symmetry and if the x-ray wavelength is small compared to all the absorption-edge wavelengths, Eq. (1)

$$f = \int_0^\infty 4\pi r^2 \rho(r) \frac{\sin kr}{kr} dr \tag{1}$$

holds. Here $k = 4\pi(\sin\theta)/\lambda$ and $\rho(r)$ is the electron density (electrons per unit volume).

A crystalline structure is one in which a unit of structure called the unit cell repeats at regular intervals in three dimensions. The repetition in space is determined by three noncoplanar vectors $\mathbf{a}_1,\mathbf{a}_2,\mathbf{a}_3$, called the crystal axes. The positions of the atoms in the unit cell are expressed by a set of base vectors $\mathbf{r}_n$. The position of atom n in the unit cell $q_1q_2q_3$ is given by Eq. (2).

$$\mathbf{R}_{nq} = q_1\mathbf{a}_1 + q_2\mathbf{a}_2 + q_3\mathbf{a}_3 + \mathbf{r}_n \tag{2}$$

See CRYSTAL STRUCTURE; CRYSTALLOGRAPHY.

For a crystal containing $N_1N_2N_3$ repetitions in the $\mathbf{a}_1\mathbf{a}_2\mathbf{a}_3$ directions, the intensity of unmodified scattering is given by Eq. (3). Here I_e is the intensity, at

$$\begin{aligned} I &= I_e \sum_{nq} f_n \exp\left[\frac{2\pi i}{\lambda}(\mathbf{s}-\mathbf{s}_0)\times\mathbf{R}_{nq}\right] \\ &\quad \times \sum_{n'q'} f_{n'} \exp\left[\frac{-2\pi i}{\lambda}(\mathbf{s}-\mathbf{s}_0)\cdot\mathbf{R}_{n'q'}\right] \\ &= I_e FF^* \frac{\sin^2\left[(\pi/\lambda)(\mathbf{s}-\mathbf{s}_0)\cdot N_1\mathbf{a}_1\right]}{\sin^2\left[(\pi/\lambda)(\mathbf{s}-\mathbf{s}_0)\cdot\mathbf{a}_1\right]} \\ &\quad \times \frac{\sin^2\left[(\pi/\lambda)(\mathbf{s}-\mathbf{s}_0)\cdot N_2\mathbf{a}_2\right]}{\sin^2\left[(\pi/\lambda)(\mathbf{s}-\mathbf{s}_0)\cdot\mathbf{a}_2\right]} \\ &\quad \times \frac{\sin^2\left[(\pi/\lambda)(\mathbf{s}-\mathbf{s}_0)\cdot N_3\mathbf{a}_3\right]}{\sin^2\left[(\pi/\lambda)(\mathbf{s}-\mathbf{s}_0)\cdot\mathbf{a}_3\right]} \end{aligned} \tag{3}$$

a distance R and angle 2θ, scattered by a free electron according to classical theory, and $FF* = |F|^2$. For an unpolarized primary beam of intensity I_o, Eq. (4)

$$I_e = I_o \frac{e^4}{m^2c^4R^2}\left(\frac{1+\cos^2 2\theta}{2}\right) \tag{4}$$

holds, and $e^4/(m^2c^4) = 7.94 \times 10^{-26}$ cm^2 if R is expressed in centimeters. Here m is the mass of the electron, c is the velocity of light, and F is the structure factor, a complex quantity given by a summation over all the atoms of the unit cell as in Eq. (5), where

$$F = \sum_n f_n \exp\left[\frac{2\pi i}{\lambda}(\mathbf{s}-\mathbf{s}_0)\cdot\mathbf{r}_n\right] \tag{5}$$

$\mathbf{s}_0$ and $\mathbf{s}$ are unit vectors in the directions of the primary and diffracted beams.

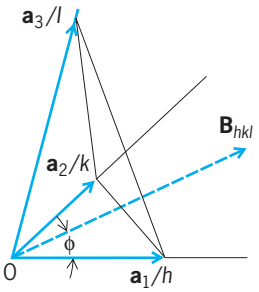


Fig. 1. Crystallographic planes with Miller indices *hkl*.

Laue equations and Bragg's law. The condition for a crystalline reflection is that the three quotients of Eq. (3) exhibit maxima, and this occurs if all three denominators vanish. Expressing the denominators in terms of three integers α, β, and γ, the three Laue equations (6) are obtained. These express the condition for a diffracted beam.

$$\begin{aligned}(\mathbf{s}-\mathbf{s}_0)\cdot\mathbf{a}_1 &= \alpha\lambda\\ (\mathbf{s}-\mathbf{s}_0)\cdot\mathbf{a}_2 &= \beta\lambda\\ (\mathbf{s}-\mathbf{s}_0)\cdot\mathbf{a}_3 &= \gamma\lambda\end{aligned} \tag{6}$$

It is convenient to introduce the concept of sets of crystallographic planes. As illustrated by **Fig. 1**, the set of planes with Miller indices *hkl* is a set of parallel equidistant planes, one of which passes through the origin, and the next nearest makes intercepts $\mathbf{a}_1/h$, $\mathbf{a}_2/k$, $\mathbf{a}_3/l$ on the three crystallographic axes.

In terms of sets of planes *hkl*, the diffraction conditions are expressed by the Bragg law, Eq. (7), where θ

$$\lambda = 2d_{hkl}\sin\theta \tag{7}$$

is the angle which the primary and diffracted beams make with the planes *hkl* and d_{hkl} is the spacing of the set. As seen from **Fig. 2**, the Bragg law is simply the condition that the path difference for rays diffracted from two successive *hkl* planes be one wavelength. In the early days of x-ray diffraction, the Bragg law was written $n\lambda = 2d \sin\theta$, and $n = 1$, 2, 3 corresponded to first-, second-, and third-order diffraction from the planes of spacing d. That notation has been largely dropped, and instead of being called second-order diffraction from planes *hkl*, it is called diffraction from the planes $2h$, $2k$, $2l$. For an extended discussion of the Bragg law *see* X-RAY POWDER METHODS.

Reciprocal lattice. The understanding and interpretation of x-ray diffraction in crystals is greatly facilitated by the concept of a reciprocal lattice. In terms of the crystal axes $\mathbf{a}_1\mathbf{a}_2\mathbf{a}_3$, three reciprocal vectors are defined by Eqs. (8). From these definitions it follows that Eq. (9) is valid. In terms of integers *hkl*, the terminal points of the vectors in Eq. (10) generate a

$$\mathbf{b}_1 = \frac{\mathbf{a}_2\times\mathbf{a}_3}{a_1\cdot\mathbf{a}_2\times\mathbf{a}_3} \qquad \mathbf{b}_2 = \frac{\mathbf{a}_3\times\mathbf{a}_1}{a_1\cdot\mathbf{a}_2\times\mathbf{a}_3} \tag{8}$$

$$\mathbf{b}_3 = \frac{\mathbf{a}_1\times\mathbf{a}_2}{\mathbf{a}_1\cdot\mathbf{a}_2\times\mathbf{a}_3}$$

$$\mathbf{a}_i\cdot\mathbf{b}_j = \begin{cases}1 & i=j\\ 0 & i\neq j\end{cases} \tag{9}$$

$$\mathbf{B}_{hkl} = h\mathbf{b}_1 + k\mathbf{b}_2 + l\mathbf{b}_3 \tag{10}$$

lattice of points called the reciprocal lattice. Each point in the lattice is specified by the integers *hkl*, and the vectors $\mathbf{B}_{hkl}$ represent two important properties of the sets of *hkl* planes: (1) $\mathbf{B}_{hkl}$ is perpendicular to the *hkl* planes, and (2) $|\mathbf{B}_{hkl}| = 1/d_{hkl}$. These two relations are readily proved from the geometry of Fig. 1. As seen, $\mathbf{a}_2/k - \mathbf{a}_1/h$ and $\mathbf{a}_3/l - \mathbf{a}_2/k$ are vectors lying in the *hkl* plane. From Eqs. (9) and (10), Eqs. (11)

$$\begin{aligned}\left(\frac{\mathbf{a}_2}{k}-\frac{\mathbf{a}_1}{h}\right)\cdot\mathbf{B}_{hkl} &= 0\\ \left(\frac{\mathbf{a}_3}{l}-\frac{\mathbf{a}_2}{k}\right)\cdot\mathbf{B}_{hkl} &= 0\end{aligned} \tag{11}$$

are obtained, and hence $\mathbf{B}_{hkl}$ is perpendicular to the planes *hkl*. The spacing of the planes *hkl* is given by Eq. (12).

$$d_{hkl} = \left|\frac{\mathbf{a}_1}{h}\right|\cos\phi = \frac{\mathbf{a}_1}{h}\cdot\frac{\mathbf{B}_{hkl}}{|\mathbf{B}_{hkl}|} = \frac{1}{|\mathbf{B}_{hkl}|} \tag{12}$$

Equivalence of the three Laue equations and the Bragg law can be shown as follows: Any vector $\mathbf{r}$ can be expressed by Eq. (13). Let $\mathbf{r}$ be the vector

$$\mathbf{r} = (\mathbf{r}\cdot\mathbf{a}_1)\,\mathbf{b}_1 + (\mathbf{r}\cdot\mathbf{a}_2)\,\mathbf{b}_2 + (\mathbf{r}\cdot\mathbf{a}_3)\mathbf{b}_3 \tag{13}$$

$(\mathbf{s}-\mathbf{s}_0)$ and combine it with the three Laue equations and Eq. (13) to obtain Eq. (14). The Bragg law

$$\mathbf{s}-\mathbf{s}_0 = \lambda\,(\alpha\mathbf{b}_1 + \beta\mathbf{b}_2 + \gamma\mathbf{b}_3) \tag{14}$$

can be written in vector form as Eq. (15) since the

$$\mathbf{s}-\mathbf{s}_0 = \lambda\mathbf{B}_{hkl} = \lambda(h\mathbf{b}_1 + k\mathbf{b}_2 + l\mathbf{b}_3) \tag{15}$$

usual form of the Bragg law is simply an equality in the magnitudes of the vectors: $|\mathbf{s}-\mathbf{s}_0| = 2\sin\theta$ and $|\mathbf{B}_{hkl}| = 1/k_{hkl}$. Comparison of Eqs. (14) and (15) shows that the integers α, β, γ of the three Laue equations are simply the Miller indices *hkl* of the Bragg law.

The positions of the atoms in the unit cell are represented by a set of atomic coordinates x_n, y_n, z_n

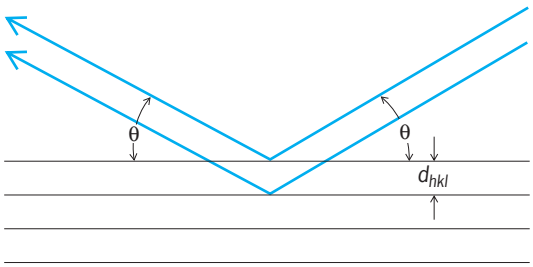


Fig. 2. Interference conditions involved in Bragg law.

such that for atom n Eq. (16) holds. For a Bragg law

$$\mathbf{r}_n = \mathbf{a}_1 x_n + \mathbf{a}_2 y_n + \mathbf{a}_3 z_n \tag{16}$$

reflection hkl, the structure factor takes the simple form of Eq. (17).

$$F_{hkl} = \sum_n f_n \exp\left[2\pi i \left(hx_n + ky_n + lz_n\right)\right] \tag{17}$$

Integrated intensity. In general, the intensity of a Bragg reflection, as expressed by Eq. (3), is not an experimentally measurable quantity. Other factors, such as the degree of mosaic structure in the crystal and the degree of parallelism of the primary beam, have a profound influence on the measured diffracted intensity for any setting of the crystal. To obtain measurements characteristic of the crystalline structure, it is necessary to adopt a more useful concept, the integrated intensity. For a small single crystal, it is postulated that the crystal is to be turned at constant angular velocity ω through the Bragg law position, and that the total diffracted energy of the reflection is to be measured. The integrated intensity E is then given by Eq. (18), where $d\alpha$ is a change in

$$E = \int\int I \frac{d\alpha}{\omega} dA \tag{18}$$

orientation of the crystal and dA is an element of area at the point of observation.

Most of the equations used in x-ray diffraction studies are derived on the assumption that the intensity of the diffracted beam is so small that any interaction with the primary beam can be neglected. These are classed as the equations for the ideally imperfect crystal. For powder samples, in which the individual crystals are extremely small, and for highly deformed single crystals, if the intensity of the diffracted beam is small, the ideally imperfect crystal is usually a good approximation. For the ideally perfect crystal, it is necessary to use a more elaborate theory which allows for the interaction of diffracted radiation with the primary beam. In general, it is the integrated intensity which is measured, and theory shows that the integrated intensity for an ideally imperfect crystal is larger than that for the ideally perfect crystal. Many of the crystalline samples used for x-ray diffraction studies are not ideally imperfect, and the measured integrated intensity is accordingly less than that predicted by the ideally imperfect crystal formulas which are used in the interpretation. The situation is usually handled by adding a correction factor called the extinction correction to the formulas for the integrated intensity from the ideally imperfect crystal.

Atomic coordinates. To have complete information about a crystalline structure, it is necessary to know all the atomic coordinates $x_n y_n z_n$ of the n atoms making up the unit cell. The atomic coordinates appear in the structure factor as given by Eq. (17), and sometimes the coordinates are obtained directly from structure factor values. Another way is to plot the electron density in the unit cell and infer the atomic positions from peaks in the electron density function. The electron density in the unit cell is given by the triple Fourier series shown in Eq. (19)

$$\rho(xyz) = \frac{1}{V} \sum_h \sum_k \sum_l F_{hkl} \times \exp\left[-2\pi i \left(\frac{hz}{a} + \frac{ky}{b} + \frac{lz}{c}\right)\right] \tag{19}$$

for which the coefficients are simply the structure factors F_{hkl}. However, from experimental measurements of either an intensity or an integrated intensity, values for $|f_{\mathrm{hkl}}|^2$ can be obtained. These yield the magnitude of F_{hkl} but not the phase. This is the most serious limitation to a straightforward determination of crystalline structures by x-ray diffraction methods. The ambiguity in the phase of F_{hkl} prevents the use of the Fourier plot of Eq. (19) as a general method for determining any crystalline structure.

Simple structures are uniquely determined by combining the x-ray intensity results with space group theory. The space group of a crystal is the repeating spatial arrangement of symmetry elements which the structure displays. Considering all the possible symmetry elements which can exist in a crystalline structure, group theory shows that there are only 230 essentially different possible combinations, and these constitute the 230 space groups. A knowledge of the macroscopic symmetry of the crystal, coupled with the systematically vanishing x-ray reflections, usually determines the space group. The limitations imposed by the space group on the possible atomic positions, coupled with the limitations imposed by the measured $|F_{hkl}|^2$, often allow a complete and unique structure determination for not too complicated structures. For highly complex structures, it is never certain that x-ray diffraction analysis can yield a complete structure determination. Additional techniques such as the isomorphous replacement by heavy atoms, the use of Patterson plots, and the determination of phase relations from inequalities are used with success on some of the complex structures. *See* GROUP THEORY.

Many structures of interest in solid-state chemistry exhibit various kinds of randomness and imperfections. The precise nature of these is sometimes of more interest than the ideal average structure. Randomness and imperfections in a structure show themselves by producing a diffuse intensity in addition to the sharp Bragg reflections. The temperature vibration of the atoms produces a diffuse intensity called temperature diffuse scattering. Quantitative measurements of this scattering lead to values for the velocity of high-frequency elastic waves and to a complete experimental determination of the spectrum of the elastic waves which constitute the thermal vibrations of the crystal. In alloys showing order-disorder changes, the short-range order parameters are obtained from quantitative measurement of the diffuse intensity which results from randomness in the atomic arrangement.

Crystalline Diffraction

The techniques employed in the study of crystalline substances are discussed below.

Laue method. The Laue pattern uses polychromatic x-rays provided by the continuous spectrum from an x-ray tube operated at 35–50 kV. The transmission Laue pattern is obtained by passing a finely collimated beam through a thin single crystal and recording the diffracted beams on a photographic film placed several centimeters beyond the crystal. For each set of planes *hkl*, θ is fixed, and the Bragg law is satisfied by selecting the proper λ from the primary beam. In a Laue pattern, the different diffracted beams have different wavelengths, and their directions are determined solely by the orientations of the *hkl* planes. Transmission Laue patterns were once used for structure determinations, but their many disadvantages have made them practically obsolete.

On the other hand, the back-reflection Laue pattern is used a great deal in the study of the orientation of crystals. The back-reflection Laue camera is shown schematically in **Fig. 3**. The polychromatic beam enters through a hole in the x-ray film and falls on a single crystal whose orientation can be set as desired by a system of goniometer circles. Diffracted beams bent through angles 2θ approaching 180° are registered on the photographic film. For cubic crystals it is very easy to read the crystal orientation from a back-reflection Laue pattern, and the patterns find considerable use in the cutting of single-crystal metal ingots.

Rotating crystal method. The original rotating crystal method was employed in the Bragg spectrometer. A sufficiently monochromatic beam, of wavelength of the order of 0.1 nanometer, is obtained by using the strong $K\alpha_1\alpha_2$ doublet with a filter which suppresses the $K\beta$ line and much of the continuous spectrum. The beam is collimated by a system of slits and then falls on the large extended face of a single crystal as shown by **Fig. 4**. Originally the diffracted beam was measured with an ionization chamber, but Geiger counters and proportional counters have largely replaced the ionization chamber. Both the crystal and the chamber turn about the spectrometer axis.

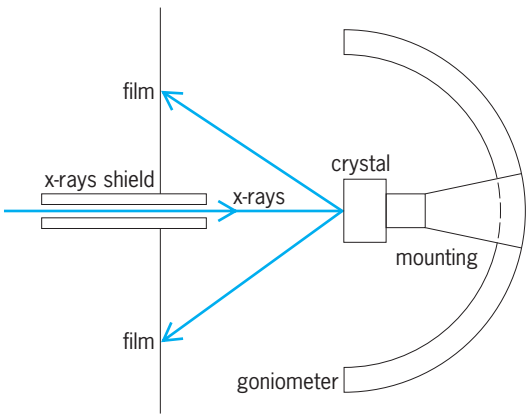


Fig. 3. Schematic of back-reflection Laue camera.

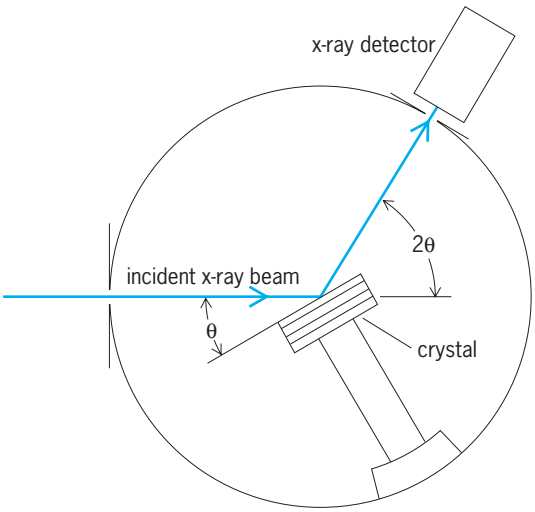


Fig. 4. Schematic of Bragg spectrometer.

The Bragg spectrometer has been used extensively in obtaining quantitative measurements of the integrated intensity from planes parallel to the face of the crystal. The chamber is set at the correct 2θ-angle with a slit so wide that all of the radiation reflected from the crystal can enter and be measured. The crystal is turned at constant angular speed ω through the Bragg law position, and the total diffracted energy E received by the ionization chamber during this process is measured. Similar readings with the chamber set on either side of the peak give a background correction. For this type of measurement, the integrated intensity E is given by Eq. (20), where P_0 is the power

$$E = \frac{P_0}{2\mu\omega}\frac{e^4}{m^2c^4}\frac{\lambda^3F^2}{v^2} \times \left(\frac{1+\cos^2 2\theta}{2\sin 2\theta}\right)\exp[-2M] \quad (20)$$

of the primary beam, μ is the linear absorption coefficient in the crystal, ω is the angular velocity of the crystal, λ is the x-ray wavelength, F is the structure factor, v is the volume of the unit cell, and $\exp[-2M]$ is the Debye factor allowing for temperature vibration. When more than one kind of atom is present, this factor must be incorporated in F^2.

Measurements of the integrated intensity E give quantitative values of F^2 directly. When a Geiger counter is used in place of an ionization chamber for this type of measurement, it is necessary to employ a narrow counterslit and traverse the counter through the reflected beam since the sensitivity of a Geiger counter is not constant over a large window opening.

The rotation camera, which is frequently used for structure determinations, is illustrated in **Fig. 5**. The monochromatic primary beam $\mathbf{s}_0$ falls on a small single crystal at O. The crystal is mounted with one of its axes (say, $\mathbf{a}_3$) vertical, and it rotates with constant velocity about the vertical axis during the exposure. The various diffracted beams are registered on a cylindrical film concentric with the axis of rotation. For a rotation about $\mathbf{a}_3$ it follows that $\mathbf{s}_0 \cdot \mathbf{a}_3 = 0$, and

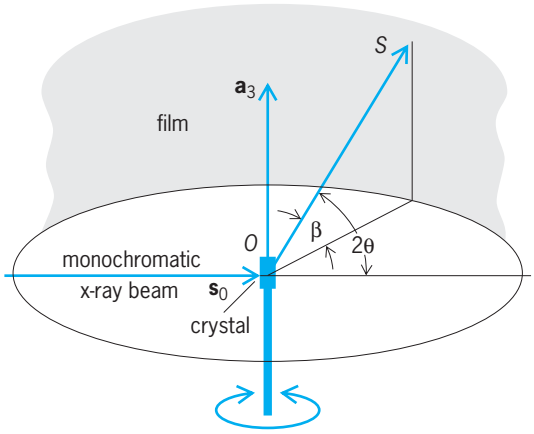


Fig. 5. Schematic of rotation camera.

the third Laue equation gives Eq. (21). The diffracted

$$\sin \beta = \frac{l\lambda}{|\mathbf{a}_3|} \tag{21}$$

beams form the elements of a set of cones, and the intersection of these cones with the cylindrical film gives a set of horizontal lines of diffraction spots. This type of pattern is called a rotation pattern, and the horizontal rows of spots are called layer lines. As seen from Eq. (21), the measured values of sin β give directly the length of the axis about which the crystal was rotated, and the layer line in which a spot occurs gives the l index of the reflection. Similar rotations about the other two axes give corresponding information. More elaborate variations of the rotation method, such as those of the Weissenberg and the precession cameras, involve a motion of the film in addition to the rotation of the crystal.

Powder method. The powder method involves the diffraction of a collimated monochromatic beam from a sample containing an enormous number of tiny crystals having random orientation. Since about 1950 an increasing number of powder patterns studies have been made with Geiger counter, or proportional counter, diffractometers. The apparatus is shown schematically in **Fig. 6**. X-rays diverging from

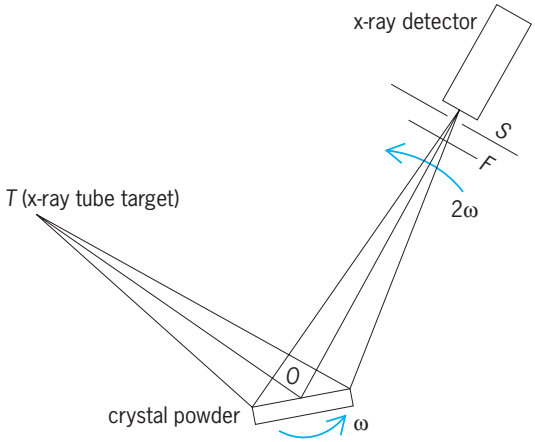


Fig. 6. Schematic representation of the Geiger counter diffractometer for powder samples.

a target at T fall on the sample at O, the sample being a flat-faced briquet of powder. Diffracted radiation from the sample passes through the receiving slit at s and enters the Geiger counter. During the operation the sample turns at angular velocity ω and the counter at 2ω. The distances TO and OS are made equal to satisfy approximate focusing conditions. A filter F before the receiving slit gives the effect of a sufficiently monochromatic beam. A chart recording of the amplified output of the Geiger counter gives directly a plot of intensity versus scattering angle 2θ.

Noncrystalline Diffraction

For a noncrystalline substance such as a glass or a liquid, a more general expression for the intensity of diffracted radiation is required. If the instantaneous position of each atom in the sample is represented by a vector $\mathbf{r}_n$, the diffracted intensity is given by Eq. (22). A particularly useful variation of Eq. (22), given as Eq. (23), is obtained by computing the average intensity $\langle I \rangle$ when the sample as a rigid array is allowed to take with equal probability all orientations in space. In Eq. (23), $k = 4\pi(\sin\theta)/\lambda$ and $r_{qn} = |\mathbf{r}_q - \mathbf{r}_n|$.

$$I = I_e \sum_q \sum_n f_q f_n \times \exp\left[\frac{2\pi i}{\lambda}(\mathbf{s} - \mathbf{s}_0) \cdot (\mathbf{r}_q - \mathbf{r}_n)\right] \tag{22}$$

$$\langle I \rangle = I_e \sum_q \sum_n f_q f_n \frac{\sin(kr_{qn})}{kr_{qn}} \tag{23}$$

The fact that there are fairly definite nearest-neighbor and second-neighbor distances in a glass or liquid means that Eq. (23) will show peaks and dips when the intensity is plotted against $(\sin\theta)/\lambda$. Peaks and dips in an x-ray diffraction pattern merely indicate the existence of preferred interatomic distances, not that the material is necessarily crystalline. X-ray patterns of noncrystalline materials are usually analyzed by a Fourier inversion of Eq. (23), which yields a radial distribution function giving the probability of finding neighboring atoms at any distance from an average atom. Bertram E. Warren

Gases. Gases and liquids are found to give rise to x-ray diffraction patterns characterized by one or more halos or interference rings which are usually somewhat diffuse. These diffraction patterns, which are similar to those for glasses and amorphous solids, are due to interference effects depending both upon the electronic distribution of each of the individual atoms or molecules and upon their relative positions in the system.

For monatomic gases the only appreciable interference effects giving rise to a distribution of scattered intensities are those produced by the electronic distribution about each nucleus. These interference effects giving rise to so-called coherent intensities are the result of the interference of the individual waves scattered by electrons in different parts of the atom. The electronic distribution of an atom is described

in terms of a characteristic atomic scattering factor which is defined as the ratio of the resultant amplitude scattered by an atom to the amplitude that a free electron would scatter under the same conditions. At zero-angle scattering the atomic scattering factor is equal to the atomic number of the atom. The coherent intensity in a given direction is proportional to the square of the atomic scattering factor. If it is assumed that the electronic distribution is spherically symmetrical, the atomic scattering factors can be readily obtained from the observed intensities. For molecular gases the interference effects depend not only on the scattering factors of the atoms but also on their relative positions in the molecule. One can observe only an average intensity scattered over a period of time during which the molecules have taken innumerable positions with respect to the incident beam. Interference effects due to the relative packing of the atoms or molecules can be neglected for dilute gases but not for dense gases.

As in the case of x-ray diffraction by crystals, light atoms such as hydrogen are difficult to detect in the presence of heavy atoms. Because of the shorter exposure times required, electron diffraction rather than x-ray diffraction has been used to study the structures of gaseous molecules. Both methods seem to be comparable in view of the accuracy of the intensity measurements and the technical difficulties involved.

Liquids. One cannot, as in the cases of dilute gases and crystalline solids, derive unambiguous, detailed descriptions of liquid structures from diffraction data. Nevertheless, diffraction studies of liquids do provide most useful information. Instead of comparing the experimental intensity distributions with theoretical distributions computed for various models, the experimental results are usually provided in the form of a radial distribution function which specifies the density of atoms or electrons as a function of the radial distance from any reference atom or electron in the system without any prior assumptions about the structure. From the radial distribution function one can obtain (1) the average interatomic distances most frequently occurring in the structure corresponding to the positions of the first, second, and possibly third nearest neighbors; (2) the distribution of distances; and (3) the average coordination number for each interatomic distance. The interpretation of these diffraction patterns given by the radial distribution function usually is not straightforward, and in general it can be said only that a certain assumed structural model and arrangement is not inconsistent with the observed diffraction data. The models considered represent only a description of the time-average environment about any given atom or molecule within the liquid.

There are great experimental difficulties in obtaining accurate intensity data. The sources of error are many for a detailed treatment the reader is referred to publications of C. Finback. A brief description of some results obtained by x-ray diffraction of liquids is given below.

Liquid elements. The radial distribution function, first used in a study of liquid mercury, has been applied to a considerable number of liquid elements mainly to compare their physical properties in the liquid and crystalline states. In most cases a lower first mean coordination number is found in the liquid state; exceptions are liquid gallium, bismuth, germanium, and lithium. The radial distribution curves give direct evidence for the existence of molecules in some liquid elements (for example, N_2, O_2 Cl_2, and P_4) and imply the existence of more complicated atomic aggregates in a few cases. Argon and helium have been extensively studied in the liquid and vapor states over wide ranges of temperature and pressure.

Liquid water and solutions. A prime example illustrating the considerable structural information made available from modern x-ray liquid diffractometry investigations is the detailed analysis of liquid water, which revealed the following significant features: (1) there are distinct structural deviations of water molecules from a uniform distribution of distances to about 0.8 nm at room temperature; (2) the first prominent maximum, corresponding to near-neighbor interactions, shifts gradually from 0.282 nm at 39.2°F (4°C) to 0.294 nm at 392°F (200°C); (3) the average coordination number in liquid water from 39.2 to 392°F (4 to 200°C) is approximately constant and slightly larger than four; and (4) the radial distribution of oxygen atoms in water at 39.2°F (4°C) is not significantly different from that in deuterium oxide at the same temperature. Comparison of calculated radial distribution functions for various proposed liquid water models (which are sufficiently defined at the molecular level) based on those derived from patterns of liquid water have shown that the only realistic model which gives agreement with data from both large- and small-angle x-ray scattering is related to a modification of the ordinary hexagonal ice structure. This solid-state structure is similar to that of the hexagonal form of silicon dioxide, tridymite, with each oxygen atom tetrahedrally surrounded by neighboring oxygen atoms to give layers of puckered six-membered rings with dodecahedral cavities large enough (radius 0.295 nm) to accommodate a water molecule.

In terms of an average configuration, the liquid water phase may be regarded as a "mixture" model comprising network water molecules forming a slightly expanded ordinary ice structure (each oxygen atom forming nearly four hydrogen bonds with neighboring oxygen atoms) and the cavity water molecules interacting with the network by less specific but by no means negligible forces. It must be emphasized that both kinds of water molecules instantaneously exist in environments which are distorted from the average, as implied by sizable root-mean-square variations in interatomic distance. *See* HYDROGEN BOND; WATER.

Radial distribution curves for concentrated $FeCl_3$ solutions indicate a large degree of local ordering of the ions with formation of Fe^{3+}—Cl^- complexes. Studies on metal-metal solutions, colloidal solutions, and molecular solutions have been made. More definite results have been obtained for concentrated solutions of strongly scattering solutes in weakly scattering solvents. Examples are the proof of the existence of a polymeric species in aqueous $Bi(ClO_4)_3$,

evidence that in aqueous solution the HgX_4^{2-} anions (X = Cl, Br, and I) are tetrahedral, and definite evidence of ion-pair formation in aqueous BaI_2.

Molten salts. A molten salt is considered to be a loose and expanded imitation of the solid with the same coordination scheme and short-range order. Careful x-ray diffraction studies of a number of molten salts have indicated that melts do not possess such quasi-crystalline structures but instead have quite open structures with a wide variety of individual ion coordinations. Interpretations of radial distribution functions for several other molten salts have been made. Liquid $AlCl_3$ appears to consist mainly of Al_2Cl_6 molecules; liquid SnI_4 is composed of independent tetrahedral molecules. The results for other molten salts are not as conclusive. *See* ELECTRON DIFFRACTION; NEUTRON DIFFRACTION.

Lawrence F. Dahl

Bibliography. C. S. Barrett and T. B. Massalski, *Structure of Metals*, 3d ed., 1980; M. J. Buerger, *Crystal Structure Analysis*, 1960, reprint 1980; B. D. Cullity, *Elements of X-ray Diffraction*, 3d ed., 2001; D. W. Hukins, *X-ray Diffraction by Disordered and Ordered Systems,* 1981; G. H. Stout and L. H. Jensen, *X-ray Structure Determination*, 2d ed., 1989.

X-ray fluorescence analysis

A nondestructive physical method used for chemical elemental analysis of materials in the solid or liquid state. The specimen is irradiated by photons or charged particles of sufficient energy to cause its elements to emit (fluoresce) their characteristic x-ray line spectra. The detection system allows the determination of the energies of the emitted lines and their intensities. Elements in the specimen are identified by their spectral line energies or wavelengths for qualitative analysis, and the intensities are related to their concentrations for quantitative analysis. Computers are widely used in this field, both for automated data collection and for reducing the x-ray data to weight-percent and atomic-percent chemical composition or area-related mass (of films). *See* FLUORESCENCE.

The materials to be analyzed may be solids, powders, liquids, or thin foils and films. The crystalline state normally has no effect on the analysis, nor has the state of chemical bonding, except for very light elements. All elements above atomic number 12 can be routinely analyzed in a concentration range from 0.1 to 100 wt %. Special techniques are required for the analysis of elements with lower atomic numbers (4–11) or of lower concentrations, and for trace analysis. The counting times required for analysis range from a few seconds to several minutes per element, depending upon specimen characteristics and required accuracy; but they may be much longer for trace analysis and thin films. The results are in good agreement with wet chemical and other methods of analysis. The method is generally nondestructive for most inorganic materials in that a suitably prepared specimen is not altered by the analytical process.

Basis of method. The theory of the method has its origin in the classic work by H. G. J. Moseley, who in 1913 measured x-ray wavelengths of a series of elements. He found that each element had a simple x-ray spectrum and characteristic wavelengths, and that there was a linear relationship between $1/\sqrt{\lambda}$ and Z, where λ is the x-ray wavelength and Z is the atomic number of the element emitting the x-ray. For example, a plot of Moseley's law can be used to show the K and L x-ray lines (**Fig. 1**). Aside from the discovery of the element hafnium in zirconium ores by G. von Hevesy, only a few practical uses of the relationship were reported until about 1950, when the introduction of modern x-ray equipment made it feasible to use x-rays for routine spectrochemical analysis of a large variety of materials.

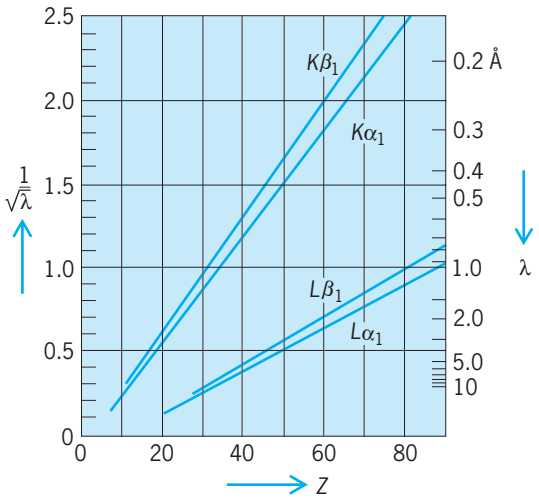


Fig. 1. Plot of Moseley's law, showing dependence of characteristic x-ray-line wavelengths λ on atomic number Z. 1 Å = 0.1 nm. (*After Philips Tech. Rev., vol. 17, no. 10, 1956*)

An x-ray source is used to irradiate the specimen, and the emitted x-ray fluorescence radiation is analyzed with a spectrometer. The fluorescence radiation is diffracted by a crystal at different angles in order to separate the wavelengths and identify the elements, and the concentrations are determined from the relative intensities. Scintillation or gas proportional counters are generally used as detectors. This procedure is widely used and is called the wavelength dispersive method.

Around 1965, lithium-drifted silicon and germanium [Si(Li) and Ge(Li)] solid-state detectors became available for x-ray analysis. These detectors have better energy resolution, and the average pulse amplitudes are directly proportional to the energies of the x-ray quanta, which can be sorted electronically with a multichannel pulse-height analyzer. This eliminates the need for the crystal and is called the energy dispersive method. Recent developments include cryogenically cooled detectors based on superconducting tunnel junctions. They combine a far better energy resolution with the ability to detect the emission lines from very light elements.

X-ray spectra. The origin of x-ray spectra may be understood from the simple Bohr model of the atom in which the electrons are arranged in orbits within the K, L, M, . . . shells. If a particle or photon

with sufficient energy is absorbed by the atom, an electron may be ejected from one of the inner shells and is promptly replaced by an electron from one of the outer shells. This results in the emission of a characteristic x-ray spectral line whose energy is equal to the difference of the binding energies of the two orbits involved in the electron transition. The new vacancy is filled by an additional transition from the outer shells, and this is repeated until the outermost vacancy is filled by a free electron. The sum of energies of all photons emitted during the vacancy-refilling cascade is the ionization energy. The energy of the emitted line from the first transition in the cascade has a slightly lower energy than the ionization energy. For example, the ionization energy for the copper K shell is 8.98 keV, and the observed lines have energies of 8.90 keV (Cu$K\beta$) and 8.04 keV (Cu$K\alpha$); the corresponding wavelengths are 0.138, 0.139, and 0.154 nanometer. Altogether the energies of x-ray *K*-lines extend over three orders of magnitude from 0.111 keV (11.2 nm, Be$K\alpha$) to 114.45 keV (0.0108 nm, U$K\beta_2$).

Optical emission lines result from resonant electron transitions in the outer (valence) shells, producing complex spectra with a large number of lines. By contrast, the x-ray lines arise only from a limited number of transitions between the high-energy levels of the inner shells, so that the x-ray spectrum of an element consists of relatively few lines. They are always initiated by a primary ionization event. Lines are named after the shell where the corresponding electron transition ends (*K*, *L*, *M*, . . . lines). The most probable transition yielding the highest line intensity in this series is named alpha, followed by beta, gamma, and others, and the indices 1, 2, 3, . . . define a specific transition within the subseries. Depending on the number of energy sublevels in each shell, there are usually only a few important lines in the *K* spectrum ($K\beta$, $K\alpha_1$, $K\alpha_2$) and a dozen or more lines in the *L* spectrum. The *M* lines are rarely used in x-ray analysis.

Auger effect. Occasionally, instead of the emission of the characteristic photon in the course of an electron transition, inner atomic absorption occurs (internal conversion or the Auger effect) when the photon appears to ionize the atom in an additional shell. The existence of an intermediate fluorescent photon is, however, denied by the quantum-mechanical explanation of the Auger effect and should serve only as an aid to illustrate the energy transfer. The ejected Auger electron has a well-defined energy, namely, the energy of the internally absorbed (virtual) photon minus its ionization energy, and can be used for chemical analysis. The probability that no Auger effect occurs, that is, that the photon is actually emitted from the atom and can be used for analysis, is called fluorescence yield. It is thereby the complementary probability to the Auger effect and is higher than 50% for K-shell ionization of elements with atomic numbers above 31 (gallium). For low-atomic-number elements, the Auger effect dominates and the fluorescence is low. This is one of the main reasons for the difficulties in the analysis of very light elements, such as berillium, boron, and carbon, where the fluorescent yield is only 10^{-4} to 10^{-3}. *See* AUGER EFFECT; ELECTRON SPECTROSCOPY.

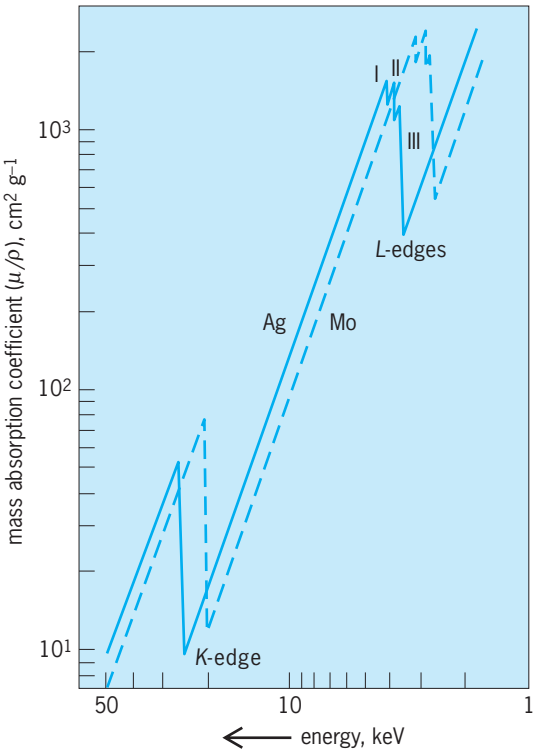


Fig. 2. Mass absorption coefficients of molybdenum (Mo) and silver (Ag) in the 1–50-keV region. Roman numerals indicate edges associated with subshells of the L shell.

X-ray absorption. The type of absorption of the photon or particle leading to the original ionization of the atom is called photoabsorption, to distinguish it from absorption by coherent scattering or Compton scattering. The probability of photoabsorption decreases gradually with increasing photon (or particle) energy, but abruptly increases by an order of magnitude when the photon energy exceeds the ionization energy of a shell. This energy is also called the absorption-edge energy (shown for the *K* and *L* edges of molybdenum and silver in **Fig. 2**). Thus the x-rays with energies just higher than the absorption-edge energy are most efficient in generating x-ray fluorescence. The efficiency decreases as the photon energy *E* is further increased from the edge approximately as $1/E^3$ or λ^3. Photons with smaller energies than the absorption edge have no effect in exciting fluorescence.

The absorption of x-rays is usually given as a mass absorption coefficient μ/ρ (usually expressed in cm^2 g^{-1}) and is independent of the physical state of the material. If more than one element is present, the weighted average of the coefficients of the individual elements is used. Tables of mass absorption coefficients have been compiled. The decrease of intensity of x-rays as they traverse the material is given by the linear absorption coefficient μ (usually expressed in cm^{-1}), obtained by multiplying the mass absorption coefficient by the density ρ of the material. The intensity decreases to $e^{-\mu x}$ of its original

value when the x-rays pass through a layer x centimeters thick.

Radiation sources. There are two general methods for producing x-ray spectra for fluorescence analysis excitation by photons and excitation by charged particles. The most common method is to expose the specimen to the entire spectrum emitted from a standard x-ray tube. It is sometimes modified by using a secondary target material (or monochromator) outside the x-ray tube to excite fluorescence. This has the advantage of selecting the most efficient energy close to the absorption edge of the element to be analyzed and reducing or not exciting other interfering elements, but the intensity is reduced by two or three orders of magnitude. Further alternatives are radioactive sources and synchrotron radiation.

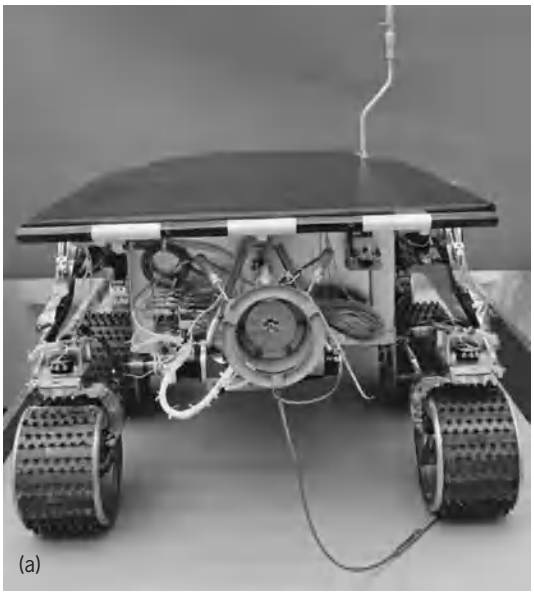
The other method, used in electron microscopes and the electron microprobe, uses an electron beam directly on the specimen, and each element generates its own x-ray spectrum, under electron bombardment, as in an x-ray tube. *See* ELECTRON MICROSCOPE.

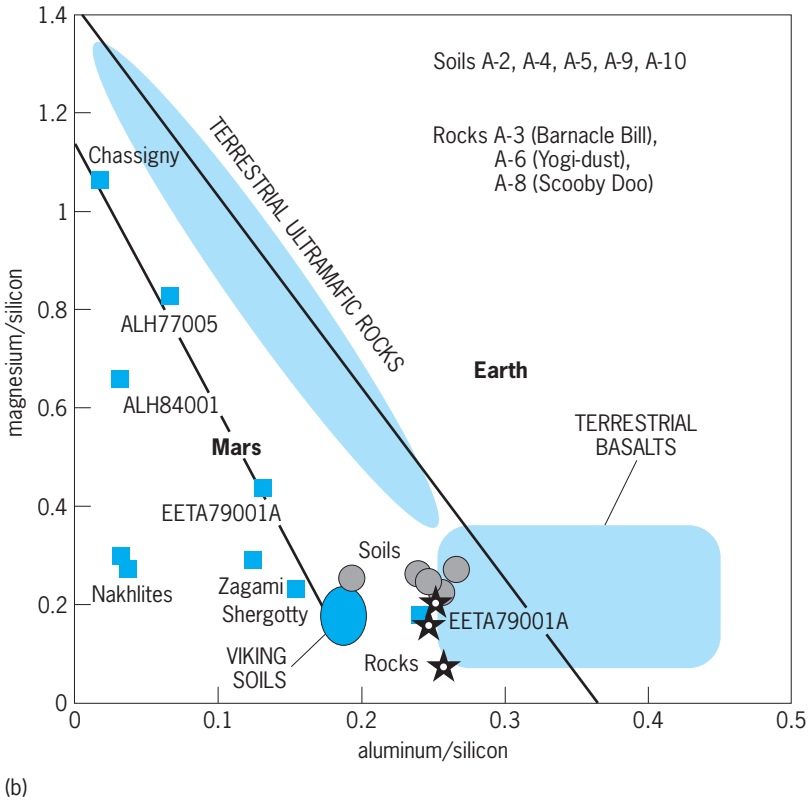
X-ray tubes. The radiative spectrum from an x-ray tube consists of continuous radiation (bremsstrahlung) and characteristic lines. Continuous radiation is emitted in the course of scattering (that is, deceleration) of electrons by the nuclei of the target atoms. Characteristic radiation is excited by electrons similarly to excitement by photons, and comes from the electronic shells. The primary x-ray-tube targets are usually tungsten, copper, rhodium, molybdenum, silver, and chromium. It is usually necessary to avoid the use of a tube whose target is identical to that of an element in the specimen, because the line spectrum from the target is scattered through the system, adding to the element signal. It is also desirable to select a target whose characteristic line energies lie closely above the absorption edges of the elements to be analyzed. For example, the WL lines and CuK lines are more efficient in exciting fluorescence in the transition elements chromium to copper than are the MoK lines; RhL lines are most useful to excite K lines of elements below sulfur in the periodic table. Tubes for fluorescence analysis usually have a single thin beryllium window placed at the side of the tube. *See* BREMSSTRAHLUNG.

Equipment is normally operated at x-ray-tube voltages of 20–60 kV in dc operation at up to 3 kW or more with water cooling. These voltages generate the K spectra of all the elements up to the rare earths and the L spectra of the higher-atomic-number elements. Since the detector is moved from point to point, it is essential to have a constant primary intensity and to stabilize the voltage and tube current. *See* X-RAY TUBE.

Radioactive isotopes. Radioactive isotopes that produce x-rays, such as iron-55 (MnK x-rays) and americium-241 (NpL x-rays), are used in place of an x-ray tube to excite fluorescence in some applications. These sources are much weaker than x-ray tubes and must be placed close to the specimen. They are often used in field applications where portability and size may be considerations. Alpha particles have been occasionally used. An example is the excitation source in the α-proton x-ray spectrometer (APXS) built into the Mars exploration vehicle *Sojourner* (Mars *Pathfinder* mission 1997/1998; **Fig. 3**). *See* RADIOACTIVITY.

Synchrotron radiation. Synchrotron radiation has many potential advantages. The continuous radiation is





(b)

Fig. 3. Alpha-Proton X-ray Spectrometer (APXS) used on Mars *Pathfinder* mission of 1997/1998. (*a*) Mars rover *Sojourner*, rear view showing spectrometer (*copyright © 1997, Jet Propulsion Laboratory, California Institute of Technology, and the National Aeronautics and Space Administration*). (*b*) Comparison of chemical composition of rocks on Earth, of various meteorites found on Earth but presumably originating from Mars, and materials analyzed by APXS near the landing site on Mars.

several orders of magnitude more intense than that of x-ray tubes and can be used with a crystal spectrometer. In addition, a tunable crystal monochromator can be placed in the incident beam to select the optimum wavelength for fluorescing each element in the specimen. Because of its high intensity and parallelism, a very narrow beam of synchrotron radiation can be masked out in order to illuminate individual spots or grains of inhomogeneous materials. Another application is ultra-trace analysis. *See* SYNCHROTRON RADIATION.

Crystal spectrometer. A single-crystal plate is used to separate the various wavelengths emitted by the specimen. Diffraction from the crystal occurs according to Bragg's law, Eq. (1), where n is a small

$$n\lambda = 2d \sin\theta \tag{1}$$

integer giving the order of reflection, λ the wavelength, d the spacing of the particular set of lattice planes of the crystal that are properly oriented to reflect, and θ the angle between those lattice planes and the incident ray. *See* X-RAY CRYSTALLOGRAPHY.

Reflection for a particular λ and d occurs only at an angle 2θ with respect to the incident ray, and it is therefore necessary to maintain the correct angular relationship of the crystal planes at one-half the detector angle. This is done by the goniometer, which is geared to rotate the crystal at one-half the angular speed of the counter tube, and therefore both are always in the correct position to receive the various wavelengths emitted by the specimen (**Fig. 4**). For a given d, there is only one angle (for each order of reflection) at which each wavelength is reflected, the angle increasing with increasing wavelength. The identification of elements by the reflection angles for their emission lines is greatly simplified by modern computer-controlled spectrometers. The angular separation of the lines, or the dispersion, given by Eq. (2), increases with decreasing d. It is thus

$$\frac{d\theta}{d\lambda} = \frac{n}{2d\cos\theta} \tag{2}$$

easy to increase the dispersion simply by selecting a crystal with a smaller d. Reducing d also limits the maximum wavelength that can be measured since $\lambda = 2d$ at $2\theta = 180^\circ$; the maximum 2θ angle that can be reached in practice with the goniometer is about 150°.

Soller slits. The crystals are usually mosaic, and the reflection is spread over a small angular range. To increase the resolution, that is, decrease the line breadth, it is necessary to limit the angular range over which a wavelength is recorded. Parallel or Soller slits are used for this purpose (Fig. 4). These slits consist of thin (0.002-in. or 0.05-mm) equally spaced flat foils of materials such as nickel and iron, and the angular aperture is determined by the length and spacing. A typical set for fine collimation would have 0.005-in. (0.13-mm) spacings and 4-in. (100-mm) length with angular aperture 0.15° and cross section 0.28 in. (7.11 mm) square. Wider angular apertures of up to a few degrees are used with multilayer mirrors for light-element analysis. The absorption of the foils is sufficiently high to prevent rays that are inclined by more than the angular aperture to extend beyond the specimen area and enter the counter tube. Two sets of parallel slits may be used, one set between the specimen and crystal and the other between crystal and detector. This greatly increases the resolution and peak-to-background ratio, and causes a relatively small loss of peak intensity.

Diffracting crystals. Crystals commonly used in spectrometers are lithium fluoride (LiF) with reflecting plane (200) or (220), silicon (111) and (220),

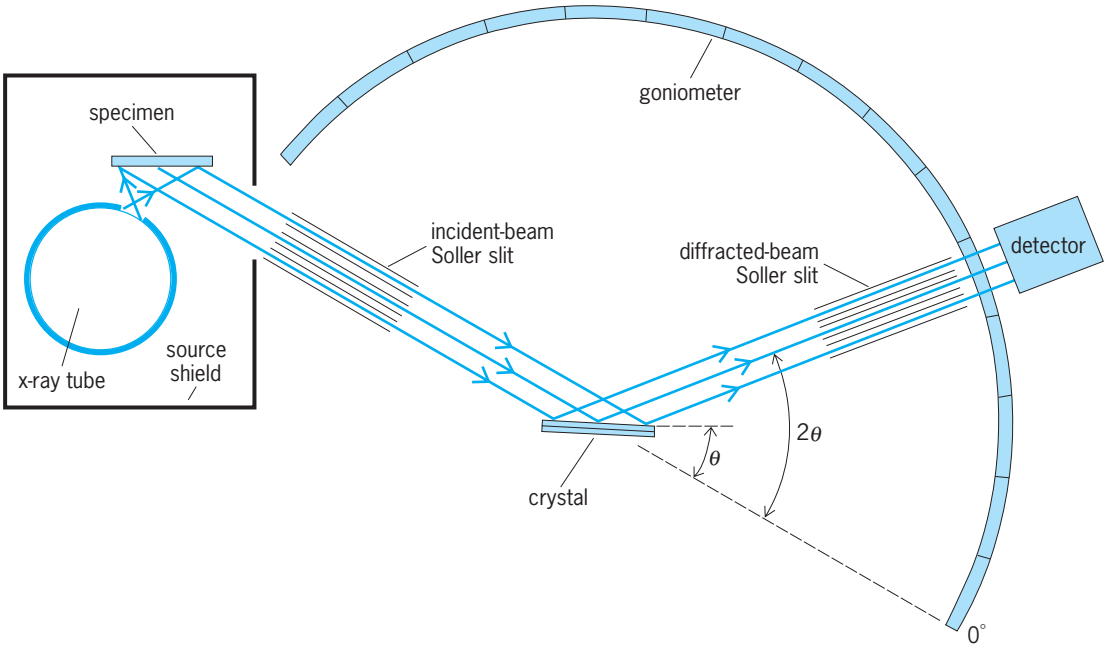


Fig. 4. X-ray fluorescence spectrograph (not to scale). Diffracted-beam Soller slit is optional.

pentaerythritol (001), acid phthalates of potassium and thallium (001), and ethylene diamine *d*-tartrate (020). It is essential that the crystal be of good quality to obtain sharp, symmetrical reflections. Unless the crystal is homogeneous, the reflection may be distorted, and portions of the reflections may occur at slightly different angles. Such effects would decrease the peak intensities of the wavelengths by varying amounts, causing errors in the analysis.

Multilayer mirrors. The longest wavelength that can be routinely analyzed with a natural crystal is around 2.4 nm (O$K\alpha$). Multilayer structures are employed as dispersive devices for lighter elements. They consist of a periodic stack of layer pairs alternating a heavy element (with high scattering power for x-rays) and light elements (serving as a spacer). The scattered partial waves from the heavy-element layers interfere constructively at certain angles in a way similar to that in crystals, but can have much longer wavelengths corresponding to the layer spacing.

Rapid analysis systems. In certain industrial applications such as the manufacture of cement, steels, and glass, and in geological exploration, large numbers of specimens containing up to a dozen or more elements must be rapidly analyzed. In some cases, the analysis must be done in a few minutes to correct the composition of a furnace that is standing by. Generally the same qualitative compositions have to be routinely analyzed, and instead of sequentially scanning over the wavelength regions, a number (up to 30) of fixed crystals and detectors are positioned around the specimen in order to allow simultaneous measurements of several elements at peak and background positions. Automated trays load the specimens into the spectrometer.

Detectors. The detectors generally used in crystal spectrometers are scintillation counters with thin beryllium windows and thallium-activated sodium iodide [NaI(Tl)] crystals for higher energies (above 4 keV), and gas flow counters with very low absorbing windows and argon/methane gas for the low-energy region (below 6 keV). A single-channel pulse-amplitude analyzer limits photon counting to a selected energy interval to improve the peak-to-background ratio and to eliminate higher-order reflections. However, no sharp energy separation is possible due to the rather limited energy resolution of these detectors. *See* GAMMA-RAY DETECTORS; PARTICLE DETECTOR; SCINTILLATION COUNTER.

Energy dispersive systems. Solid-state detectors with good energy resolution are used in conjunction with a multichannel pulse-amplitude analyzer. No crystals are required, and the detector and specimen are stationary during the measurement. The method is used with either electron-beam excitation in electron microscopes or with x-ray-tube sources. The photons of various energies are registered, and their energies are determined as soon as they enter the detector. As this occurs statistically for the various fluorescence line energies, the acquisition of the spectral data appears to be simultaneous for all lines.

Solid-state detectors. Lithium-drifted silicon [Si(Li)] detectors are generally used for the lower energies of fluorescence analysis, while lithium-drifted germanium [Ge(Li)] detectors are more often used for nuclear high-energy gamma-ray detection. The energy resolution of good Si(Li) detectors is below 130 eV (full width at one-half maximum) for Mn$K\alpha$ radiation. The lithium-drifted detectors require cooling during operation, for which liquid nitrogen is often used. *See* JUNCTION DETECTOR.

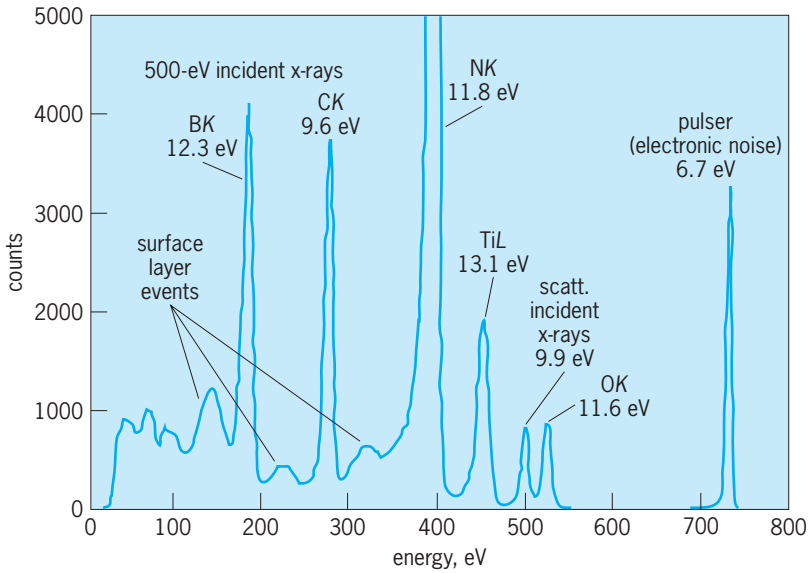


Fig. 5. Spectrum of boron nitride partially covered with titanium powder obtained with a cryogenically cooled superconducting tunnel junction detector. The energy resolution of all lines up to several hundred electronvolts is around 10–12 eV. A crystal spectrometer with a multilayer mirror would have a resolution of about 16 eV at B$K\alpha$. (*After M. Frank et al., Cryogenic high-resolution x-ray spectrometers for SR-XRF and microanalysis, J. Synchrotron Rad., 5:515–517, 1998*)

The resolution of the detector is closely linked to its temperature. Some types allow operation at room temperature with degraded resolution, or with Peltier cooling stages. The most recent development are superconducting tunneling junction devices, which are operated at liquid helium temperature. Their energy resolution is comparable to wavelength dispersive spectrometers or is even much better, particularly for light elements (**Fig. 5**).

Analyzer. The output signals from the detector are fed into the analyzer, where the photon counts are stored in memory locations (1024–8192 channels are generally used) that are related to the energies of these photons. This also allows visual observation on a cathode-ray-tube screen of the accumulated spectrum and of the simultaneous counting process. Analyzers are usually provided with cursor markers to easily identify the peaks in the spectrum. Computer memories can be used for storage of the spectral counts, thus providing efficient access to computer routines for further data evaluation.

Use. Energy dispersive x-ray spectrometers are useful to accumulate spectra in short time intervals (for example, 1 min) that often allow a preliminary interpretation of the qualitative and quantitative composition of the specimen. The instruments are comparatively small, because they are designed to accept a large aperture of radiation. They require only low-power x-ray tubes that sometimes can be air-cooled.

Limitations. An important limitation of energy dispersive systems with Si(Li) detectors is the energy resolution, which is about an order of magnitude poorer in the lower energy region than that of crystal spectrometers. For example, the $K\alpha$ lines of the transition elements overlap with the $K\beta$ lines of the element preceding it in atomic number, causing severe analytical difficulties in an important region of the spectrum. The peak-to-background ratio is significantly lower than in crystal spectrometers because of the lower resolution. Another limitation is that the maximum number of photons that can be processed by the electronic circuits is limited to about 15,000–50,000 counts per second. This is the total photon count from the entire detected spectral region. Trace elements with low count rates in a matrix of high-count elements are therefore difficult to detect with sufficient statistical accuracy. Various attempts have been made to overcome this drawback by selectively exciting the elements of interest by using selective filters or secondary targets, which also greatly reduces the amount of x-ray-tube radiation that is scattered into the detector.

Microanalysis. The electron microprobe is widely used for elemental analysis of small areas. An electron beam of 1 micrometer (or smaller) is used, and the x-ray spectrum is analyzed with a focusing (curved) crystal spectrometer or with an energy dispersive solid-state detector. Usually two or three spectrometers are used to cover different spectral regions. Light elements down to beryllium, boron, and carbon can be detected. An important use of the method is in point-to-point analysis with a few cubic micrometers of spatial resolution. *X-Y* plots of any element can be made by moving the specimen to determine the elemental distribution.

Figure 6 illustrates the spectra obtained with three of the most frequently used methods of analysis. The specimen, a high-temperature alloy of the type used in aerospace and other industries, was prepared by the National Institute of Standards and Technology with stated composition in weight percent: molybdenum (Mo) 3.13, niobium (Nb) 4.98, nickel (Ni) 51.5, cobalt (Co) 0.76, iron (Fe) 19.8, chromium (Cr) 17.4, titanium (Ti) 0.85, and aluminum (Al) 0.085, total 99.27%.

Figure 6*a* shows the high-resolution spectrum obtained in about an hour with a lithium fluoride (LiF; 200) crystal spectrometer using 50-kV, 12-milliampere x-ray-tube excitation and scintillation counter. This spectrum also contains the second-order (II) and third-order (III) crystal reflections of molybdenum and niobium whose $K\beta_1$ and $K\beta_3$ components are resolved. The lower resolution of the energy dispersive method is shown in Fig. 6*b*, recorded in about 10 min using 50-kV, 2-microampere x-ray-tube excitation, Si(Li) detector, and 40 eV per channel (about 400 channels are shown). The spectral range includes the unresolved molybdenum and niobium *L* lines and titanium. Figure 6*c* is an energy dispersive spectrum excited by a 25-keV electron beam. The molybdenum and niobium spectra are weakly excited at this low voltage and are not visible on the scale used in the plot. The differences in the relative intensities of the lines in the spectra arise from differences in the conditions of excitation and detection, and they illustrate the necessity of using the proper correction factors for each method of analysis to derive the correct weight percent composition.

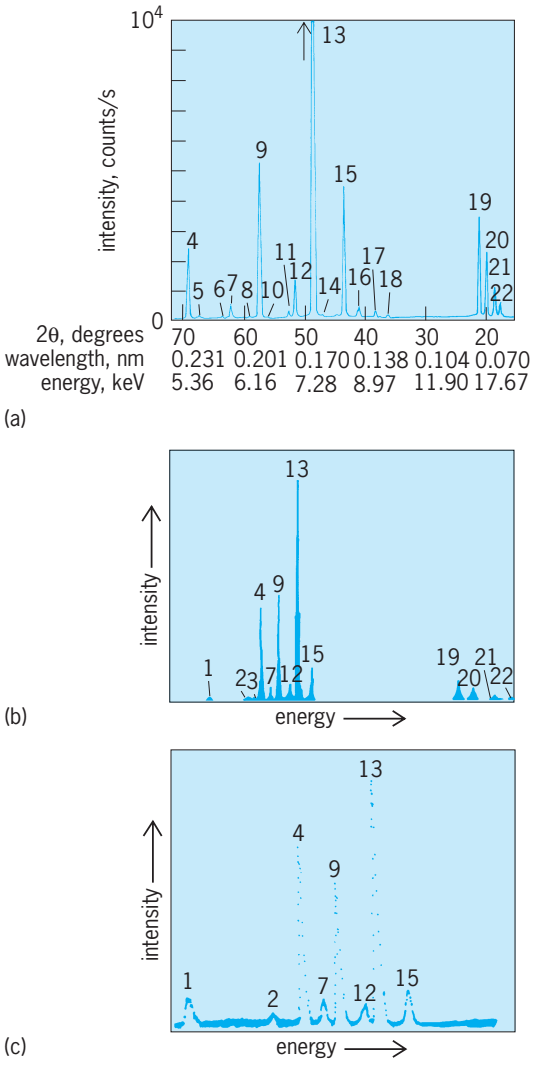


Fig. 6. Fluorescence spectra of high-temperature alloy obtained with (*a*) crystal spectrometer, (*b*) energy dispersive method with x-ray-tube excitation, and (*c*) energy dispersive method with electron-beam excitation. Spectral lines: 1, Mo + Nb$L\alpha$ + $L\beta$. 2, Ti$K\alpha$. 3, Ti$K\beta$. 4, Cr$K\alpha$. 5, Nb$K\alpha_{1,2}$III. 6, Mo$K\alpha_{1,2}$III. 7, Cr$K\beta$. 8, Nb$K\beta_{1,3}$III. 9, Fe$K\alpha$. 10, Mo$K\beta$III. 11, Co$K\alpha$. 12, Fe$K\beta$. 13, Ni$K\alpha$. 14, Co$K\beta$. 15, Ni$K\beta$. 16, Mo$K\alpha_{1,2}$II. 17, Nb$K\beta_{1,3}$II. 18, Mo$K\beta_{1,3}$. 19, Nb$K\alpha$. 20, Mo$K\alpha$. 21, Nb$K\beta_{1,3}$. 22, Mo$K\beta_{1,3}$.

Specimen preparation. The specimens may be in the form of powders, briquettes, solids, thin films, or liquids. The surface exposed to the primary x-ray beam must be flat, smooth, and representative of the sample as a whole, because usually only a thin surface layer contributes to the fluorescent beam in a highly absorbing specimen. The thickness of this layer is called information depth and may be only a micrometer or less for electron-beam excitation and 10–100 μm or more for x-rays. The degree of surface roughness, which is difficult to measure quantitatively, causes losses in intensity and results in

errors in the analysis. Consequently, solid samples are generally polished; and then, if necessary, they are lightly etched or specially cleaned to remove contaminants. This is particularly important when light elements are measured. Special care must be taken when a measured element is a constituent of such surface contamination.

Powders. Powders are processed in one of two ways. The first is to press the ground material into briquettes. The pressure should be several tons per square centimeter (1 ton/cm^2 equals approximately 15,000 lb/in.^{2} or 100 megapascals), and in most cases organic binders have to be used to improve the mechanical stability. The second way is to use fusion techniques, where the powders (mostly mineralogical or metal oxides) are dissolved at high temperatures in borax or similar chemicals, and glassy pellets are obtained after cooling. The advantage of the second method is a high homogeneity of the specimen and a reduction of interelement effects; but the intensities are reduced.

Liquids. Liquids can be analyzed by using small containers with a thin window cover. Examples are sulfur determination in oils during the refining process, lubrication oil additives, the composition of slurries, and the determination of lead, zinc, and other elements in ore processing. Low concentrations of elements in solution can be concentrated with specific ion-exchange resins and collected on filter papers for analysis. Gases containing solid particles can be filtered and the composition of the particles determined as for atmospheric aerosol filters for environmental studies. In certain industrial applications, liquids are continuously analyzed while flowing through a pipe system with a thin window in the x-ray apparatus.

Quantitative analysis. The observed fluorescent intensities must be corrected by various factors to determine the concentrations. These include the spectral distribution of the exciting radiation, absorption, fluorescence yield, and others. Two general methods have been developed to make these corrections: the fundamental parameter method and the empirical parameter method.

Fundamental parameter method. In the fundamental parameter method, a physical model of the excitation is developed and described mathematically. The method derives its name from the fact that the physical constants, like absorption coefficients and atomic transition probabilities, are also called fundamental parameters. Primary and secondary excitation are taken into account; the first is the amount of fluorescent radiation directly excited by the x-ray tube. Secondary excitation is caused by other elements in the same specimen, whose fluorescent radiation has sufficient energy to excite the characteristic radiation of the analyzed element. In practical applications, the count rate must be calibrated for each element by comparing it to the count rate from a standard of accurately predetermined composition. A standard may contain several elements or can be a pure element.

The fundamental parameter method is capable of accuracies around 1% (absolute weight percentage) for higher concentrations, and between 2 and 10% (relative) for low concentrations. The method has the advantage of allowing the use of pure-element standards. Significantly higher accuracies can be obtained with standard specimens of similar composition to the unknown.

The fundamental parameter method can also be used to determine thickness and chemical composition of thin films.

Empirical parameter method. The empirical parameter method is based upon simple mathematical approximation functions, whose coefficients (empirical parameters) are determined from the count rates and concentrations of standards. A widely used set of approximation functions is given by Eq. (3), where c_i

$$\frac{C_i}{r_i} = \frac{1 + \sum_{j \neq i}^{n} \alpha_{ij} C_j}{R_i} \qquad i = 1, \ldots, n \qquad (3)$$

is the concentration of the analyzed element i in the unknown specimen, r is the corresponding count rate, R_i is the count rate from a pure-element specimen i, C_j are the concentrations of the other elements in the unknown specimen, n is the number of elements, and α_i are the empirical parameters (also called alpha coefficients).

A minimum of $n - 1$ standard specimens, each of which contains the full set of n elements (or a correspondingly higher number, if they contain fewer elements), is required to calculate the empirical parameters, α_{ij}, before actual analysis of an unknown is possible. In practical applications, however, at least twice as many standards should be used to obtain good accuracy, thus requiring considerable effort in standard preparations. The empirical parameter method is therefore mainly used in routine applications, where large numbers of similar specimens must be analyzed. The accuracy of the method depends upon the concentration range covered by the standards; around $\pm 0.1\%$ or better can be obtained if a set of well-analyzed standards with similar compositions to the unknowns are used. If pure-element standards are not available, the pure-element counts rates, R_i in Eq. (3), can also be determined by computation from additional multielement standards.

Trace analysis. There are two distinct analytical tasks that are called trace analysis: the detection or quantification of small amounts of a material (possibly a pure element), and the determination of very low concentrations in an abundantly available sample. In both cases, the relationship between concentration and count rates is practically linear. The minimum detection limit is defined by that amount or concentration for which the peak is just statistically significant above background level B, usually $3B^{1/2}$. The background arising from scattered continuous radiation from the x-ray tube is a limiting factor in determining the peak-to-background ratio. Since intensity measurements can theoretically be made arbitrarily accurate by using long counting times, the

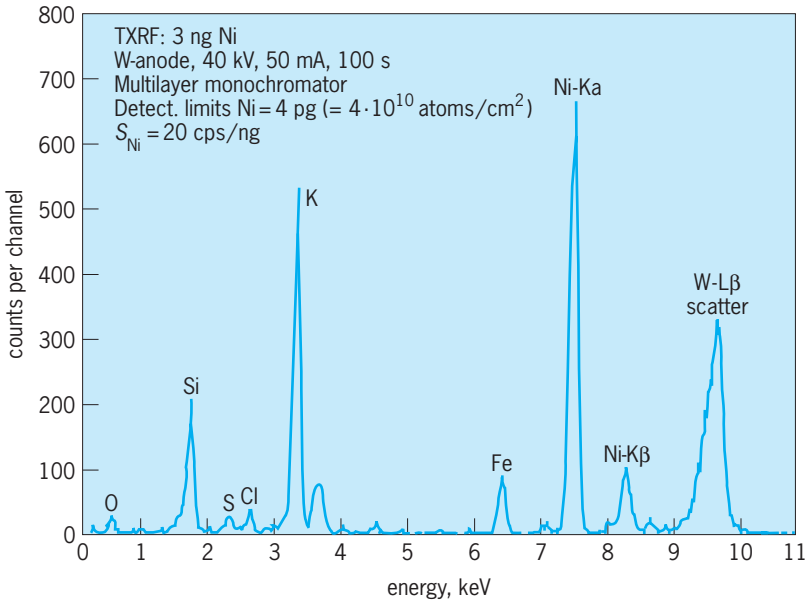


Fig. 7. Example of trace analysis by total reflection x-ray fluorescent analysis (TXRF). A droplet containing 3 ng dissolved nickel (Ni) was applied to a substrate (silicon-wafer), dried, and measured. The sensitivity *S* in this particular setup was 20 counts per second and per nanogram Ni, and the theoretical detection limit for 100 s counting time was 4 picograms, corresponding to 4×10^{10} atoms/cm^2. The elements sulfur (S), potassium (K), and iron (Fe) are contaminants of the solvent, and the silicon (Si) and oxygen lines originate mainly from a thin silicon dioxide (SiO_2) layer on top of the wafer. (*Data provided by P. Wobrauschek, Atominstitut der Österreichischen Universitäten, Vienna*).

minimum detection limits could be indefinitely low. However, in practice, the limiting factors are the background level and long-term instrument drift. Depending upon excitation conditions, matrix, and counting times, traces in the parts-per-million region may be detected with conventional instruments, and in the parts per trillion region by total reflection x-ray fluorescence.

Total reflection XRF (TRXFA). Ultra-trace analysis by x-ray fluorescence is possible by a special technique and instrumentation which is based upon background suppression by total reflection of the primary x-ray beam. The physical explanation is that the index of refraction of x-rays is very slightly smaller than 1, and a beam impinging at a flat surface at angles of a few tenths of a degree is totally reflected without noticeably penetrating the material. In practice, a substrate of a silicon-single crystal (such as a wafer) is used and a small droplet of dissolved analyte material applied and dried. The x-ray beam penetrates only the sample material, not the substrate. A Si(Li) energy dispersive detector is placed at close distance to the specimen. With conventional x-ray tubes, detection limits in the picogram range have been reported, and in the femtogram range by using synchrotron radiation. Total reflection x-ray fluorescence analysis instrumentation is commercially available (**Fig. 7**).

Thin-film analysis. As a rule of thumb, materials with a thickness exceeding a few hundred micrometers can be considered "infinitely thick" from the viewpoint of x-ray fluorescence. This limit decreases by a factor 5–20 for light elements. The intensities of thinner specimens are correspondingly lower, depending upon element, matrix, and experimental setup. In the analysis of very thin films (a few tens of nanometers), the count rates are a linear function of element concentration and of film thickness. Absorption and interelement effects must be taken into account in the analysis of thicker films and foils. This can be done with special fundamental parameter methods, but it requires adequate computing power for efficient evaluation of data.

Fundamental parameter methods allow the determination of thickness and element concentrations of thin films as well as individual layers in multilayer structures. Limitations apply to common elements of two or more layers and with respect to very light elements.

Limitations on accuracy. In both the fundamental parameter and empirical parameter methods, limitations of the accuracy are due mainly to uncertainties in the composition of the standards and variations in the specimen preparation; intensity fluctuations due to counting statistics and instrument instabilities may also contribute.

Supplemental methods. As in all analytical methods, it is sometimes necessary to supplement the chemical data from fluorescence analysis with data by other methods to properly characterize the material. The first three elements in the periodic table (hydrogen, helium, lithium) cannot be measured by x-ray fluorescence, because none of their emission lines are in the x-ray regime. The light elements beryllium through magnesium (including such important elements as carbon, oxygen, and nitrogen) can be measured, but frequently with difficulties. Often they are crucial in the characterization of a specimen, such as carbon in steels, and oxygen in rocks and oxide samples, which may require optical emission, atomic absorption, Auger and electron spectroscopy, or other analytical methods. *See* ANALYTICAL CHEMISTRY; ATOMIC SPECTROMETRY; SURFACE PHYSICS.

An important supplementary method is x-ray polycrystalline diffraction, in which the crystalline chemical phases are identified by comparing the pattern of the unknown with standard patterns. Computer methods are widely used to search the 40,000 phases currently contained in the Powder Diffraction File published by the International Center for Diffraction Data, Newtown Square, Pennsylvania. Mixtures of phases can be quantitatively determined, and there are no limitations on the chemistry of the substances. By combining the chemical data from fluorescence with the phase data from diffraction, the relation between the constituents of the sample and its properties can be established. *See* X-RAY DIFFRACTION.

Applications. X-ray fluorescence analysis is widely used for compositional control in large-scale industrial processing of metals and alloys, cements, the petroleum industry, and inorganic chemicals. Among the many other major applications are geological exploration and mineralogical analysis, soils and plants, glasses, corrosion products, the analysis of raw materials, and the measurement of plating coating thickness. It is an important method in materials characterization for research and technology,

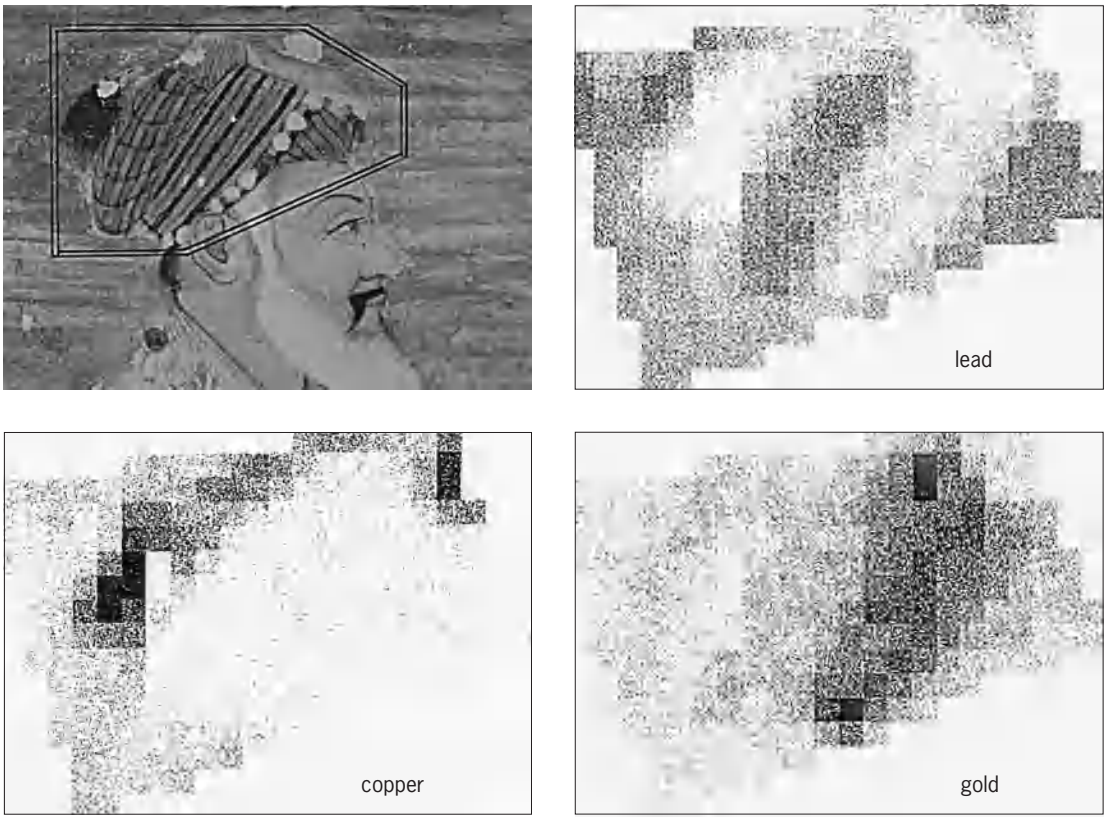


Fig. 8. Analysis of pigments in an Indian miniature, Mughal period, seventeenth century, Schloss Schönbrunn, Vienna. In the detail, the headdress (marked area) is approximately 2.1 × 1.6 in. (50 × 40 mm) and was measured in pixel steps of 0.1 in. (2.5 mm). The distribution of the elements copper, lead, and gold is shown, indicating usage of lead-white, minimum (red), azurite (blue), malachite (green), and metallic gold. (*Analysis by M. Schreiner, Akademie der Bildenden Künste, Vienna. Copyright, Österreichisches Bundesdenkmalamt, Vienna*)

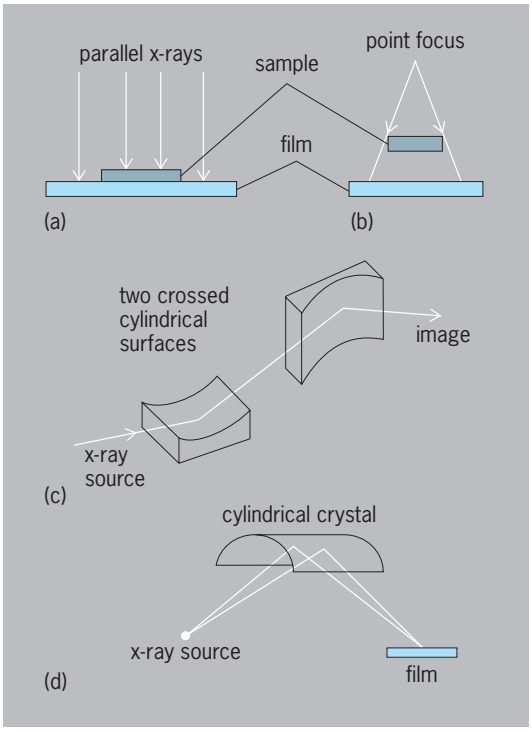
providing chemical information without destroying the sample. It is the only feasible method for many complex analyses that would require extremely long times by conventional wet chemical methods on materials such as the refractory metals, high-speed cutting steels, and complex alloys.

Besides the large-scale industrial applications, the method has been used in a variety of analyses in the medical field, for environment protection and pollution control, and for many research applications. Examples are trace analysis of heavy metals in blood; analysis of airborne particles, historic coins, potteries, lead and barium in Roman skeletons, and various elements in archeological specimens; analysis of pigments to establish authenticity of a painting (**Fig. 8**); quality control of noble metals in alumina-based exhaust catalysts for cars; and analysis of ash and sulfur in coals, slags from furnace products, and surface deposits on bulk metals. The method is also widely used in forensic problems, where it is often combined with x-ray powder diffraction. Remote analysis of rocks using x-ray spectrometers carried by spacecraft and stellar landers has proven to be a valuable source of information in search of the origin of the solar system and its planets. William Parrish; Michael Mantler

Bibliography. *Advances in X-ray Analysis*, annually; E. P. Bertin, *Introduction to X-ray Spectrometric Analysis*, 1978; K. F. J. Heinrich, *Electron Beam X-ray Microanalysis*, 1981; K. F. J. Heinrich et al. (eds.), Energy Dispersive X-ray Spectrometry, *NBS Spec. Publ.*, no. 604, 1981; R. Jenkins, *X-ray Fluorescence Spectrometry*, 1988; R. Jenkins, R. W. Gould, and D. Gedke, *Quantitative X-ray Spectrometry*, 2d ed., 1995; G. R. Lachance, F. Claisse, and H. Chassin, *Quantitative X-ray Fluorescence Analysis: Theory and Application*, 1995; K. L. Williams, *Introduction to X-ray Spectrometry*, 1987.

X-ray microscope

A term used to describe a technique and an instrument or combination of instruments which utilize x-radiation for chemical analysis and for magnification of 100–1000 diameters. The resolution possible is about 0.25 micrometer. X-ray microscopy is a relatively recent development among the microscopic techniques. The contrast in the x-ray microscopic image is caused by varying x-ray attenuation in the specimen. The advantage of x-ray microscopy is that it yields quantitative chemical information, besides structural information, about objects, including those which are opaque to light. It is a reliable ultramicrochemical analytical technique by which amounts of elements and weights of samples as small as 10^{-12} to 10^{-14} g can be analyzed with an error of only a few percent. *See* X-RAY OPTICS.



Principles for x-ray microscopy. (*a*) Contact microradiography. (*b*) Projection x-ray microscopy. (*c*) Reflection x-ray microscopy. (*d*) X-ray image spectrography.

Principles of image formation. There are four general principles of x-ray microscopy: (1) contact microradiography (**illus.** *a*), (2) projection x-ray microscopy (illus. *b*), (3) reflection x-ray microscopy (illus. *c*), and (4) x-ray image spectrography (illus. *d*). For a discussion of the first two *see* MICRORADIOGRAPHY

Contact microradiography. In contact microradiography the thin specimen is placed in close contact with an extremely fine-grained photographic emulsion which has a resolution of more than 1000 lines/mm, and radiographed with x-rays of suitable wavelength. Thus an absorption image in scale 1:1 is obtained, and this image is subsequently viewed in a light microscope. The maximal resolution is that of the optical microscope (0.25 μm), but the image has more information, which can be obtained by examining the microradiogram in the electron microscope. The optical microscope gives more information because it is possible to correlate the intensity of light in the image with the primary attenuation in the specimen. This is not at present possible in the electron microscope. The electron microscope gives only high magnification, but it is not possible to draw conclusions about the quantitative attenuation of electrons in the specimen in order to draw conclusions about the specimen composition. *See* ELECTRON MICROSCOPE.

Projection x-ray microscopy. Projection x-ray microscopy, or x-ray shadow microscopy, is based on the possibility of producing an extremely fine x-ray focal spot. This is achieved by an electronic lens system similar to that in the electron microscope. The fine focal spot is produced on a very thin metal foil which serves as a transmission target. The x-rays are generated on the target by the impact of the electrons. The sample is placed near the target, and the primary magnification depends on the ratios of the distances from focal spot to sample and sample to film. Resolution is of the same order as the size of the focal spot; the best value is about 0.1 μm in favorable objects.

Reflection x-ray microscopy. The method of reflection x-ray microscopy is based on the fact that the refractive index for x-rays in solids is a very small amount less than 1. Thus at grazing incidence (that is, incidence at very small angles), the x-rays are totally reflected, and if the reflecting surface is made cylindrical, there will be a focusing action in one dimension. By crossing two such surfaces a true image formation can be obtained, although with some astigmatism, which can be corrected by giving the surfaces a complicated optical shape. The resolution by this procedure is about 0.5–1 μm.

X-ray image spectrography. X-ray image spectrography utilizes Bragg reflections in a cylindrically bent crystal and produces slightly enlarged emission images; this technique is best classified as a micromodification of x-ray fluorescence analysis. The resolution is about 50 μm. *See* X-RAY DIFFRACTION; X-RAY FLUORESCENCE ANALYSIS.

Chemistry of specimen. The penetration of x-rays through material varies with the amount and composition of the attenuator (the material) and the wavelength of the x-rays. The specimen must be thin in order to obtain high resolution, and therefore soft (low voltage) and ultrasoft x-rays must be used. For biological specimens, x-rays of energies between 0.2 and 5 kV (about 5.0–0.2 nanometers in wavelength) are generally used. Thin metallurgical specimens can be examined with harder x-rays because of their higher x-ray absorption. In general the x-ray absorption increases with increasing wavelength of the x-rays (softer rays) and increasing atomic number of the elements composing the specimen. Thus a microscopic structure containing elements of high atomic number of sufficient concentration embedded in a matrix of elements with low atomic numbers will show up as a heavily x-ray-absorbing structure. At certain wavelengths in the x-ray-absorption spectrum, discontinuities appear: the so-called x-ray-absorption edges. The position of these edges in the spectrum is unique for each element; thus they can be utilized for identification of elements in microscopic structures in a specimen. *See* X-RAYS.

Quantitative microscopy. By measuring the variation of density in the x-ray microscopic image the x-ray attenuation can be calculated. If the x-ray microscopic image is recorded with x-rays of suitable wavelength (often monochromatic x-rays), certain chemical characteristics of the specimen can be quantitatively assessed. If a certain element is to be determined in a microstructure, two x-ray microscopic images are recorded with monochromatic x-rays with wavelengths

on either side and close to the absorption edge for the particular element. In this way elementary analysis can be performed on specimens weighing not more than 10^{-10} to 10^{-12} g with a relatively high degree of accuracy. Instead of using a photographic film to record the x-ray image, various types of detectors (Geiger-Müller tubes) are used to measure the variations of x-ray transmission in the sample. Such techniques, especially in the form of scanning, may become more useful in the future.

By proper selection of x-ray wavelength the dry weight, water content, and the content of certain other compounds can be determined in cellular structures down to about 1 μm in size. Thus weights as small as 10^{-14} g can be determined with an analytical error of only a few percent.

Applications. In biology, x-ray microscopy has been utilized for the quantitative determination of the dry weight, water content, and elementary composition of many tissues, for example, nerve cells, various types of secretory cells, the individual bands in muscle fibers, chromosomes, and parts thereof. Especially in studying mineralized tissues much new information has been gained through the use of x-ray microscopy.

The capillary circulation in the living animal can be studied by x-ray microscopy. Thorotrast and Umbradrie, contrast media, may be injected into the bloodstream and successive microradiograms recorded. The procedure is called microangiography and has been applied to the study of the finest blood vessels.

As indicated previously, soft and ultrasoft x-rays must be used for the x-ray microscopy of thin sections or smears of biologic tissue. These long-wavelength x-rays must be generated in specially designed x-ray tubes. The sample and photographic emulsion are enclosed within the high vacuum of the x-ray tube. The pictures obtained show the distribution of dry weight (mass) within the cells and tissues in addition to information on the structures in the specimen.

It is possible to determine the thickness of thick samples by using oblique incidence of the x-rays. With this technique, the thickness of nerve fibers and of some constituents of bone tissue has been determined.

By tilting the film and specimen at a certain angle, stereoscopic microradiograms have been obtained. When thick sections of tissue have been used, the three-dimensional arrangement of bone cells in bone tissue and the three-dimensional image of the capillary net in the circulatory system have been made available for study.

X-ray microscopy has been applied to biology, medical research, mineralogy, and metallurgy among other fields.

Arne Engstrom

Bibliography. A. J. Morgan, *X-ray Microanalysis in Electron Microscopy for Biologists,* 1985; D. Sayre et al. (eds.), *X-ray Microscopy Two*, 1988; K. Shinohara and K. Yada (eds.), *X-ray Microscopy in Biology and Medicine*, 1990.

X-ray optics

By analogy with the science of optics, those aspects of x-ray physics in which x-rays exhibit properties similar to those of light waves. X-ray optics may also be defined as the science of manipulating x-rays with instruments analogous to those used in visible-light optics. These instruments employ optical elements such as mirrors to focus and deflect x-rays, zone plates to form images, and diffraction gratings to analyze x-rays into their spectral components. X-ray optics is important in many fields, including x-ray astronomy, biology, medical research, thermonuclear fusion, and x-ray microlithography. It is essential to the construction of instruments that manipulate and analyze x-rays from synchrotrons and particle storage rings for synchrotron radiation research. *See* GEOMETRICAL OPTICS; OPTICS; PHYSICAL OPTICS; X-RAYS.

X-ray refraction and absorption. When W. C. Roentgen discovered x-rays in 1895, he unsuccessfully attempted to reflect, refract, and focus them with mirrors, prisms, and lenses of various materials. The reason for his lack of success became evident after it was established that x-rays are electromagnetic waves of very short wavelength for which the refractive index of all materials is smaller than unity by a only a small decrement. In addition, x-rays are absorbed by materials. The refractive index can be written as a complex quantity, as in Eq. (1),

$$\tilde{n} = 1 - \delta - i\beta \tag{1}$$

where $1 - \delta$ represents the real part, n, of the refractive index and β is the absorption index. These quantities are strongly dependent on the wavelength of the x-rays (see **table**) and the material. X-rays of wavelength about 0.1 nanometer or less are called hard x-rays and are relatively penetrating, while x-rays of wavelength 1–10 nm are less penetrating and are called soft x-rays. Radiation in the wavelength range 10–50 nm, called the extreme-ultraviolet (EUV) region, is very strongly absorbed by most materials. Values of δ remain very small throughout the x-ray and extreme-ultraviolet regions with the consequence that radiation is very weakly refracted by any material. Thus lenses for x-rays would have to be very strongly curved and very thick to achieve an appreciable focusing effect. However, because the absorption index, β, is so high in comparison, such thick lenses would absorb most of the incident radiation, making such lenses impractical. *See* ABSORPTION OF ELECTROMAGNETIC RADIATION; REFRACTION OF WAVES; ULTRAVIOLET RADIATION.

Dependence of the complex refractive index of copper on x-ray wavelength

	Refractive index parameter	
Wavelength, nm	δ	β
0.1	1.1×10^{-5}	8.7×10^{-7}
2.0	2.8×10^{-3}	6.0×10^{-3}
15.0	4.2×10^{-2}	6.9×10^{-2}

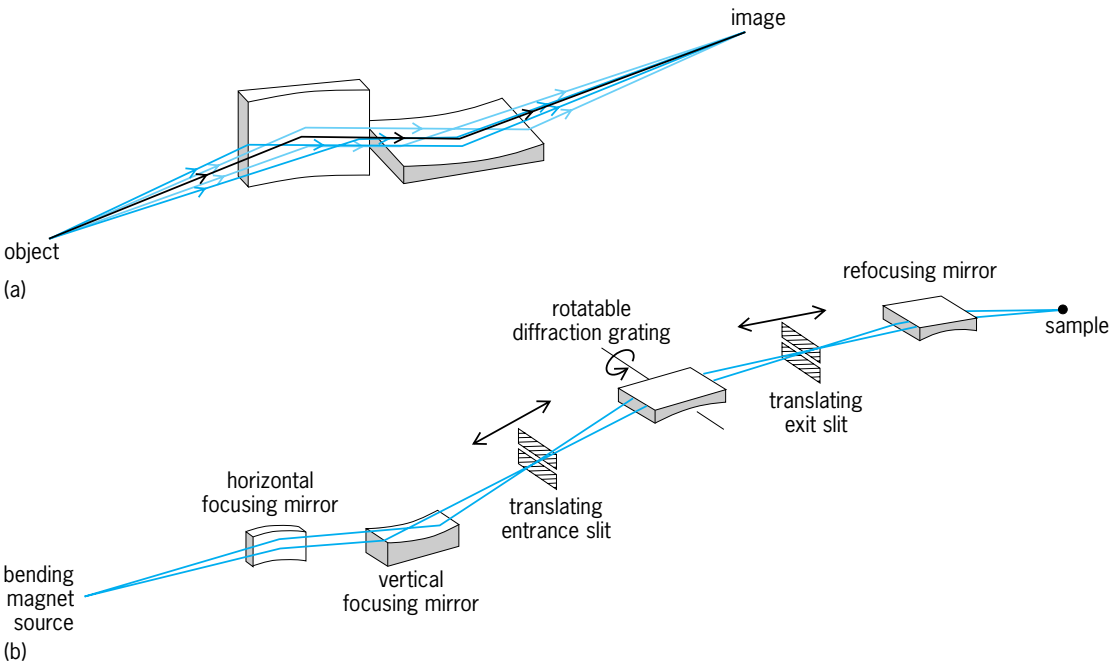


Fig. 1. Microscope principle that corrects the astigmatism associated with glancing-incidence spherical mirrors. (*a*) Refocusing of x-rays to a point source by using a combination of two spherical mirrors (*after J. H. Underwood and D. T. Attwood, The renaissance of x-ray optics, Phys. Today, 37(4):44–50, April 1984*). (*b*) Use of this principle in a synchrotron radiation beam line to collect, analyze, and refocus the x-rays emitted by the circulating electrons.

X-ray mirrors. If radiation is incident normally (that is, perpendicular) to a surface between two media of differing refractive index, the fraction of the energy that is reflected is $^1/_4(\delta^2 + \beta^2)$. This is clearly impractically small for a normal-incidence mirror for x-rays. However, useful mirrors can be constructed by using the principle of total reflection. If electromagnetic waves are incident on the boundary between one material of refractive index n_1 and another of lower refractive index n_2, there exists an angle of incidence I_c, called the critical angle, given by Eq. (2). If the angle

$$\sin I_c = \frac{n_2}{n_1} \tag{2}$$

of incidence (the angle of incident radiation with respect to the normal to the surface) is greater than this critical angle, all the wave energy is reflected back into the first medium. This phenomenon can be seen when looking upward into an aquarium tank; objects in the tank are reflected in the surface of the water, which acts as a perfect mirror. An analogous situation occurs for x-rays. Since the refractive index for all materials is slightly less than 1, x-rays incident from vacuum (or air) on a polished surface of, say, a metal encounter a lower refractive index and there exists a critical angle given by $\sin I_c = 1 - \delta$. Since δ is very small, I_c is very close to 90°. In this case the angle of incidence is customarily measured from the tangent to the surface rather than from the normal, and the angle $\theta_c = 90^\circ - I_c$ is termed the angle of glancing (or grazing) incidence. This angle is typically in the range 0.1–1.0°. *See* REFLECTION OF ELECTROMAGNETIC RADIATION.

Reflecting x-ray optics. Glancing incidence reflection of x-rays was first demonstrated in 1922. One of the first applications was the absolute measurement of the wavelength of x-rays, using a reflecting diffraction grating consisting of numerous parallel grooves of precisely known separation ruled on a reflecting surface. However, attempts to construct a reflecting system to produce x-ray images were long frustrated by the severe image defects, or aberrations, suffered by mirrors used at glancing incidence. The most severe of these is astigmatism; whereas a mirror in normal incidence forms reasonably good images of a point source, at glancing incidence these become line foci separated by a large distance. In 1948 P. Kirkpatrick and A. Baez solved this problem by crossing two spherical mirrors (**Fig. 1***a*). However, microscopy using reflection optics has never achieved wide acceptance for biological or medical research because of the difficulty of removing the remaining aberrations, and image degradation caused by scattering from mirror-surface imperfections. Such microscopes have been used to photograph the implosion of fuel pellets in laser-fusion experiments. *See* ABERRATION (OPTICS); DIFFRACTION GRATING; NUCLEAR FUSION; X-RAY MICROSCOPE.

Optics in the Kirkpatrick-Baez configuration plays an important role in constructing beam lines for research at synchrotrons and storage rings. A typical beam line consists of three to seven glancing-incidence mirrors and diffraction gratings (Fig. 1*b*). These components perform the functions of collecting the x-rays emitted by the circulating electrons, separating the x-rays into spectral components by using monochromators, and focusing the x-rays onto experimental samples. *See* SYNCHROTRON RADIATION.

Astigmatism can alternatively be corrected with a single glancing-incidence mirror having a toroidal or tubelike shape. Such a mirror must have a radius of curvature in the direction along the tube axis which is $1/\sin^2 \theta_c$ times greater than the radius in the perpendicular direction. Such mirrors are more difficult to make than the spherical or cylindrical forms used in the Kirkpatrick-Baez system. A true image-forming device requires two such reflecting elements to correct additional aberrations such as coma. Such a two-mirror device was invented by H. Wolter in 1952. The first mirror of a Wolter microscope is an ellipse of revolution, the second a confocal hyperboloid. A few microscopes of this type have been used in laser-fusion research and in synchrotron radiation microscopy. Woltier x-ray telescopes, designed to observe objects at infinity, have for the first element a paraboloid of revolution in place of an ellipsoid. They have been used extensively in x-ray astronomy aboard orbiting spacecraft. *See* X-RAY ASTRONOMY; X-RAY TELESCOPE.

Multilayer optics. Although the reflectivity of surfaces at glancing angles greater than the critical angle is very small, this reflectivity can be enhanced by depositing a stack of ultrathin films having alternately high and low values of δ on the surface (**Fig. 2**). The individual thicknesses of these films is adjusted so that the reflections from each interface add in phase at the top of the stack in exact analogy to the multilayer mirrors used for visible light. However, whereas visible multilayers require film thicknesses of hundreds of nanometers, in the x-ray region the thickness of each film must be between 1 and 100 nm. Such ultrathin films can be made by a variety of vacuum deposition methods, commonly sputtering and evaporation. The response of these artificial multilayers is strongly wavelength-selective. For incident x-rays of wavelength λ, the reflectivity of the stack has peaks at the angles θ_n given by the Bragg equation (3). Ordinarily the multilayers are used in

$$n\lambda = 2d \sin \theta_n \qquad (3)$$

the first order, that is, $n = 1$. *See* X-RAY DIFFRACTION.

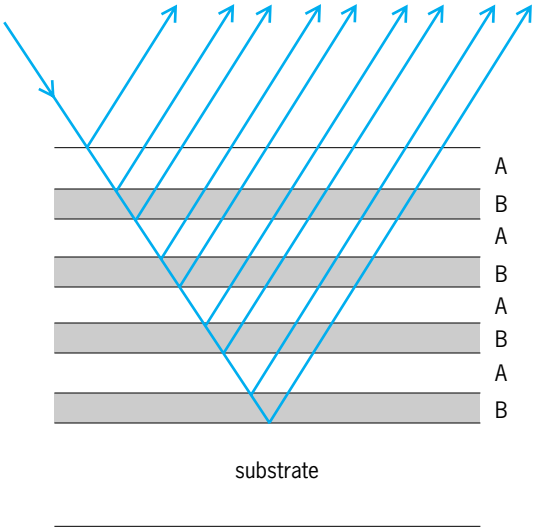


Fig. 2. Reflection of x-rays by a mirror, with alternating layers of materials A and B. The weak reflected beams from each of the interfaces combine with the correct phase relationship to form a strong reflected beam.

As a coating for glancing-incidence optics, multilayers allow a mirror to be used at a shorter wavelength (higher x-ray energy) for a given glancing angle, increasing the projected area and thus the collection efficiency of the mirror. At wavelengths longer than 3 or 4 nm, multilayer mirrors can be used to make normal-incidence mirrors of relatively high reflecting power. For example, stacks consisting of alternating layers of molybdenum and silicon can have reflectivities as high as 65% at wavelengths of 13 nm and longer. These mirrors have been used to construct optical systems that are exact analogs of mirror optics used for visible light. For example, normal-incidence x-ray telescopes have photographed the Sun's hot outer atmosphere at wavelengths of around 18 nm. Multilayer optics at a wavelength of 13.5 nm can be used to perform x-ray microlithography by the projection method to print features of dimensions less than 100 nm.

Crystal optics. Crystals are natural multilayer structures and thus can reflect x-rays. Many crystals can be bent elastically (mica, quartz, silicon) or plastically (lithium fluoride) to make x-ray focusing reflectors. These are used in devices such as x-ray spectrometers, electron-beam microprobes, and diffraction cameras to focus the radiation from a small source or specimen on a film or detector. Until the advent of image-forming optics based on mirrors and zone plates, the subject of x-ray diffraction by crystals was called x-ray optics. *See* X-RAY CRYSTALLOGRAPHY; X-RAY SPECTROMETRY.

Zone plates. Zone plates are diffraction devices that focus x-rays and form images. They are diffracting masks consisting of concentric circular zones of equal area, and are alternately transparent and opaque to x-rays. Whereas mirrors and lenses focus radiation by adjusting the phase at each point of the wavefront, zone plates act by blocking out those regions of the wavefront whose phase is more than a half-period different from that at the plate center. Thus a zone plate acts as a kind of x-ray lens. Zone-plate microscopy is the most promising candidate method for x-ray microscopy of biological specimens. *See* DIFFRACTION.

James H. Underwood

Bibliography. B. Aschenbach, X-ray telescopes, *Rep. Prog. Phys.*, 48:579–629, 1984; G. Margaritondo, *Introduction to Synchrotron Radiation*, 1988; E. Spiller, *Soft X-ray Optics*, 1994.

X-ray powder methods

Physical techniques used for the identification of substances, and for other types of analyses, principally for crystalline materials in the solid state. In these techniques, a monochromatic beam of x-rays is directed onto a polycrystalline (powder) specimen, producing a diffraction pattern that is recorded on

film or with a diffractometer. This x-ray pattern is a fundamental and unique property resulting from the atomic arrangement of the diffracting substance. Different substances have different atomic arrangements or crystal structures, and hence no two chemically distinct substances give identical diffraction patterns. Identification may be made by comparing the pattern of the unknown substance with patterns of known substances in a manner analogous to the identification of people by their fingerprints. The analytical information is different from that obtained by chemical or spectrographic analysis. X-ray identification of chemical compounds indicates the constituent elements and shows how they are combined.

The x-ray powder method is widely used in fundamental and applied research; for instance, it is used in the analysis of raw materials and finished products, in phase-diagram investigations, in following the course of solid-state chemical reactions, and in the study of minerals, ores, rocks, metals, chemicals, and many other types of material. The use of x-ray powder diffraction methods to determine the actual atomic arrangement, which has been important in the study of chemical bonds, crystal physics, and crystal chemistry, is described in related articles. *See* X-RAY CRYSTALLOGRAPHY; X-RAY DIFFRACTION.

Lattice geometry. The atoms in crystalline substances are arranged in a symmetrical three-dimensional pattern: some atomic arrangement is repeated by the symmetry of the crystal along straight lines throughout the crystal. The array of points and lines in **Fig. 1** outlines a lattice or framework of a typical crystal in which the third dimension is normal to the plane of the drawing. The smallest group of atoms which has the symmetry of the entire pattern is called the unit cell. There are several ways in which the unit cell is selected. In this case it was drawn parallel to the crystallographic axes a, b, and c. The traces of the various lattice planes (normal to the drawing) are indicated by heavy lines. The method used by crystallographers to identify these planes is as follows. For a given set of planes, count the number of planes crossed from one lattice point to the next along a, then repeat the procedure along b and c. The resulting numbers, called $hk\ell$, respectively, are known as the Miller indices of that set of planes, and assignment of indices to each line is called indexing the pattern.

The spacings d between the planes are related to the Miller indices and the unit cell dimensions. In crystals of the cubic system, the crystallographic axes are normal to each other and have the same length, $a = b = c$; and d is given by Eq. (1). Similar

$$d = \frac{a_o}{\sqrt{(h^2 + K^2 + \ell^2)}} \qquad (1)$$

relations exist for the other five crystal systems. The relations for the low-symmetry systems, monoclinic and triclinic, are much more complicated. Various types of charts to facilitate indexing tetragonal, orthorhombic, and hexagonal substances are available.

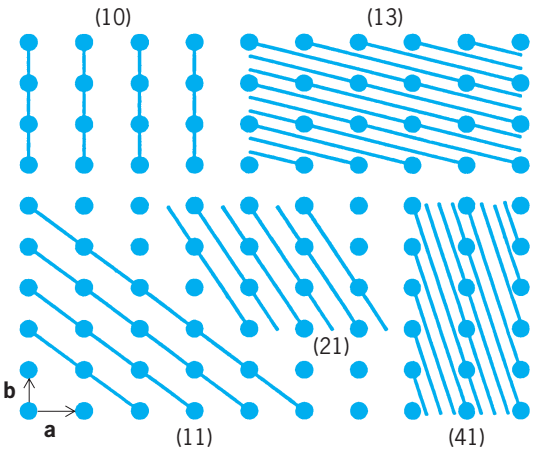


Fig. 1. Lattice of a typical crystal. The third dimension is normal to the plane of the drawing. Traces of various lattice planes are indicated by heavy lines, and their Miller indices are given.

See CRYSTAL STRUCTURE; CRYSTALLOGRAPHY.

Bragg's law. Atomic diameters have approximately the same dimensions as the wavelengths of x-rays, and therefore the crystal can act as a three-dimensional grating for x-rays in a manner analogous to the diffraction of ultraviolet or visible light by a ruled one-dimensional grating. Under appropriate conditions, the electrons around each atom scatter the incident x-ray beam in a coherent manner and in certain specific directions; the scattering from billions of atoms is in phase at the same time. *See* DIFFRACTION GRATING.

Shortly after the discovery of x-ray diffraction by M. von Laue in 1912, this complex phenomenon was formulated in a simple relation by William H. Bragg and William L. Bragg, in which the diffraction is visualized as a reflection from a large number of parallel planes. (This should not be confused with the total reflection of x-rays, which occurs at very small grazing angles from highly polished surfaces.) They showed that when two or more parallel rays of the same wavelength are incident at the same glancing angle Θ to a set of atomic planes, the path difference of the reflected rays from adjacent planes is one wavelength. This may be expressed as Eq. (2), where

$$\lambda = 2d_{hk\ell} \times \sin \Theta \qquad (2)$$

λ is the wavelength of the incident x-rays and $d_{hk\ell}$ the interplanar spacing between the $(hk\ell)$ planes of atoms. Thus the conditions for x-ray reflection are very restrictive because these is only one angle Θ at which the x-rays of a given wavelength are reflected by a particular set of atomic planes of spacing d. For a given wavelength, larger d spacings appear at smaller angles. Moreover, differentiation of the Bragg equation yields Eq. (3). This equation shows that the shift

$$\frac{\Delta\Theta}{\Delta d} = \frac{\tan \Theta}{d} \qquad (3)$$

in line position $\Delta\Theta$ due to a change of lattice spacing Δd increases as the tangent of the angle and reaches a maximum at $\Theta = 90^\circ$ (reflection angle $2\Theta = 180^\circ$),

where tan Θ is infinite. Hence the highest-angle lines in the back-reflection region of the pattern are the most sensitive to changes in the lattice spacings and consequently supply the most accurate data for the measurement of the unit cell dimensions.

Characteristics of powder patterns. Many materials are not available in the form of large single crystals, and moreover it is impractical to obtain all the x-ray reflections from single crystals for identification purposes. If the sample does not already exist in polycrystalline form, it may be pulverized. When a fine-grained powder consisting of thousands of small, randomly oriented crystallites is exposed to the x-ray beam, all the possible reflections from the various sets of atomic planes can occur simultaneously. **Figure 2** shows two examples of diffraction patterns recorded for barium titanate ($BaTiO_3$). Four basic pieces of information can be derived from the experimental pattern: (1) the 2Θ value from which the *d*-spacing can be calculated; (2) the absolute intensity, from which relative intensities can be calculated; (3) the peak width; (4) the form of the background.

For historical reasons having to do with data storage and reproduction problems, and experimental pattern is typically reduced to yield a table of *d*-values and relative intensities. These data are called the reduced pattern.

The complexity of the pattern is determined primarily by the symmetry of the substance rather than by its chemical composition. Hence, chemically complex compounds can have patterns that are nearly as simple as those from iron or copper. The number of lines also increases with unit cell size. When two or more substances are present in the sample, the pattern of each substance appears independent of the others. This makes the identifications more difficult, but mixtures containing as many as six substances have been successfully analyzed. It is also possible to make quantitative analyses of the mixtures by comparing the relative intensities of one or more principal lines of each substance. Usually some reference standards of known chemical composition are prepared to facilitate the interpretation.

One of the most important characteristics of powder patterns is that isostructural substances (that is, substances with the same crystal structure) give similar diffraction patterns. **Figure 3** shows diffraction patterns of the various phases of quartz (α-SiO_2), which have different crystal structures. Indeed, many substances may occur in two or more crystal structures; that is, they may have polymorphic forms but the same chemical composition. Such forms may be caused by slight differences in chemical preparation, different heat treatments, and other factors. *See* POLYMORPHISM (CRYSTALLOGRAPHY); QUARTZ.

When a substance is strained or plastically deformed, the x-ray lines broaden, and when the strain is removed by annealing, the lines return to their original sharpness. Figure 2*a* shows the diffractometer recording of a sample of barium titanate that has been strained by crushing, while Fig. 2*b* shows the same substance in an unstrained state. X-ray patterns can thus be used to follow the course of heat treatment or other processes used to remove strains.

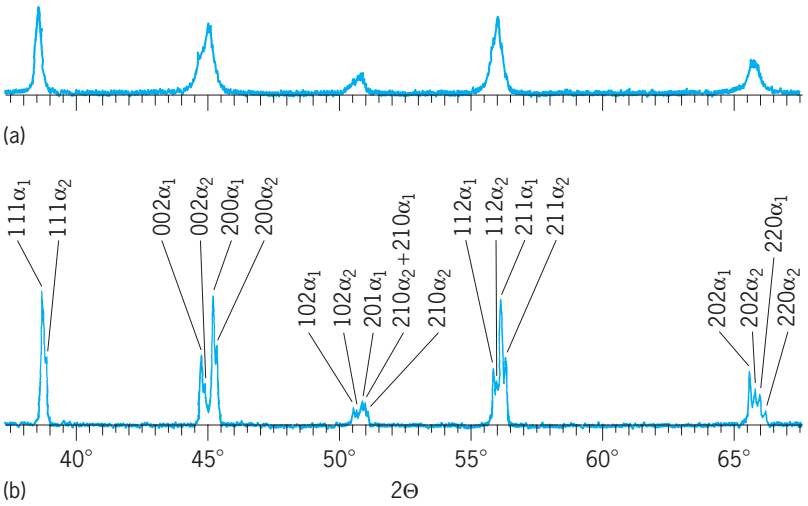


Fig. 2. Diffractograms of barium titanate ($BaTiO_3$). (*a*) Strained specimen with broadened diffraction lines. (*b*) Unstrained specimen. (*After Transactions of the Instrumentation and Measurements Conference, Stockholm, 1952*)

If the crystallites do not have a completely random orientation, the line shapes and relative intensities will change accordingly. For example, in rolling thin sheets of metal or in drawing wire, the crystallites align themselves in special ways, depending on the mechanical conditions of the process. The special or preferred orientation of the crystallites gives x-ray diffraction patterns which may be markedly different from those of the random crystallites. Similar conditions may arise in electroplating, where the plating conditions or the substrate may cause the crystallites to have a nonrandom orientation. Comparison of the random and oriented x-ray patterns shows the degree of orientation in the sample. *See* ELECTROPLATING OF METALS.

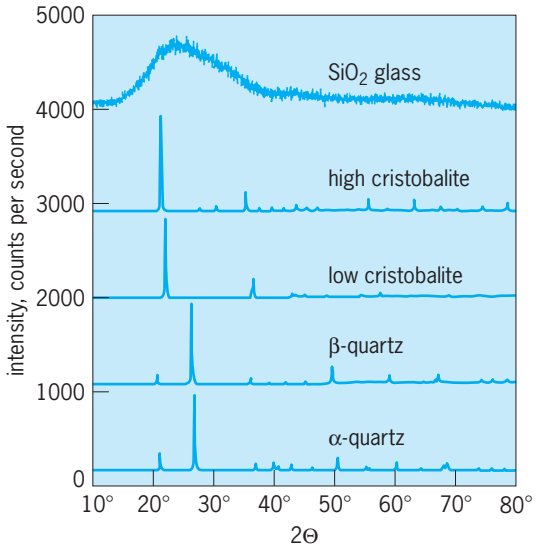


Fig. 3. Diffraction patterns of polymorphic phases of SiO_2 (quartz). (*After R. Jenkins and R. L. Snyder, Introduction to X-ray Powder Diffractometry, Wiley, 1996*)

TABLE 1. Types of x-ray diffraction databases

Name	Content	Location
Powder Diffraction File (PDF)*	Inorganic, organic, mineral, and so on	Newtown Square, PA
Cambridge Structural Database (CSD)†	Organic, organo-metallic	Cambridge, U.K.
Inorganic Crystal Structure Database (ICSD)‡	Inorganic materials	Karlsruhe, Germany
NRCC Metals Data File (CRYSTMET)	Metals and alloys	Ottawa, Canada
Protein Data Bank (PDB)	Structure of macromolecules	Brookhaven, NY
NIST Crystal Data [NBS(CD)]	Inorganic and organic unit cells	Gaithersburg, MD

*Maintained by International Centre for Diffraction Data (ICDD).
†Maintained by Cambridge Crystallographic Data Centre (CCDC).
‡Maintained by Fachinformationszentrum (FIZ).

When the crystallites are very small, for instance, about 1 micrometer, the x-ray lines broaden by an amount which increases with decreasing size. Comparisons with the line breadths of samples of larger crystallite sizes may lead to a measure of the average crystallite size in the specimen.

Amorphous materials such as glasses and liquids give patterns which consist of only a few broad lines superimposed on a continuous background. As an example, the upper diffractogram in Fig. 3 shows an amorphous pattern. Patterns of various amorphous substances closely resemble each other, and hence the method is impractical for identification. On the other hand, since the two types of patterns are so different, the method is ideally suited to distinguish between crystalline and amorphous substances and to determine the degree of crystallinity of substances between the two extremes. The study of the progress of devitrification (crystallization) of a glass and similar problems are frequently accomplished with x-rays. Certain chemical and structural properties of liquids might be ascertained from the diffraction patterns when the liquids are frozen to form crystals. *See* AMORPHOUS SOLID; GLASS; LIQUID.

There are also many smaller changes in the x-ray pattern which may reveal important information. In substitutional solid solutions, for example, atoms of different elements may substitute for one another and occupy the same relative positions as in the pure metals. The substitutions of solute atoms occur on the same lattice sites occupied by the solvent atoms, but are randomly distributed. If the atoms are of different size, the average unit cell size will change accordingly. In simple cases, it is possible to determine the chemical composition of intermediate members by measuring the unit cell dimensions because there is often a nearly linear relationship between the two. In interstitial solid solutions, atoms are added to the empty spaces in the structure and there is little, if any, change in the dimensions. *See* ALLOY STRUCTURES; SOLID SOLUTION.

Diffraction databases. There are a number of databases available for x-ray diffraction work (**Table 1**). The majority of these databases are designed and maintained for the single-crystal community rather than for the powder community. Nevertheless, much cross fertilization takes place. For example, a number of the patterns in the Powder Diffraction File (PDF) are calculated from single-crystal data of the type contained in the other databases.

While these databases have proven their usefulness in a wide range of applications, there has been little attempt to exploit a combination of them. Agreements allow mutual use of the PDF, the Cambridge Structural Database (CSD), and the Inorganic Crystal Structure Database (ICSD). Three major advantages have accrued from this cooperation. (1) By use of cross-reference "hooks" for each database entry, the user has access to experimental powder data from the PDF and structural information from the ICSD, permitting the full modeling of the experimental pattern. (2) The PDF can be supplemented using powder patterns calculated from structural data. (3) The combined efforts of the various editorial groups can only help to improve the overall content and quality of diffraction data.

In addition, the International Centre for Diffraction Data (ICDD), which maintains the PDF, archives fully digitized raw data sets (PDF-3). The availability of the trace of the original experimental data can often give useful background information for the editorial process. In addition, the success rate of the search-match process has been dramatically improved by a strategy based on searching the whole observed pattern with its background (not just the reduced pattern), and on adding candidate phases together to compose, rather than decompose, an observed multiphase pattern.

Powder Diffraction File. The Powder Diffraction File is a collection of single-phase x-ray powder diffraction patterns in the form of tables of the interplanar spacings (d) and relative intensities (I^{rel}) characteristic of the compound. New and updated diffraction patterns are continually added to the PDF, and an automated editorial system allows their detailed review. Currently 2500 such patterns are added each year, comprising 1900 inorganic patterns and 600 organic patterns. There is a continuing effort to ensure that a significant proportion of the new phases represent current needs and trends in industry and research. The master database of powder patterns is continually undergoing revision and updating, but in order to ensure that all database users have the opportunity to work with the same version, a frozen version of the master database is produced each year as the PDF-2.

The PDF-2 contains a series of individual data

33-1161 ★

SiO_2

Silicon Oxide

Quartz, syn

Rad. CuKα_1 **λ** 1.540598 **Filter** Mono. **d-sp** Diff.
Cut off **Int.** Diffractometer **$I/I_{cor.}$** 3.6
Ref. *Natl. Bur. Stand. (U.S.) Monogr. 25*, **18** 61 (1981)

Sys. Hexagonal **S.G.** $P3_221$ (154)
a 4.9133(2) **b** **c** 5.4053(4) **A** **C** 1.1001
α **β** **γ** **Z** 3 **mp**
Ref. Ibid.

D_x 2.65 **D_m** 2.66 **SS/FOM** F_{30}=77(.013,31)

εα **nωβ** 1.544 **εγ** 1.553 **Sign** + **2V**
Ref. Swanson, Fuyat, *Natl. Bur. Stand. (U.S.), Circ. 539*, **3** 24 (1954)

Color Colorless
Pattern taken at 25 C. Sample from the Glass Section at NBS, Gaithersburg, Maryland, USA, ground single-crystals of optical quality. Pattern reviewed by Holzer, J., McCarthy, G., North Dakota State University, Fargo, North Dakota, USA, *ICDD Grant-in-Aid* (1990). Agrees well with experimental and calculated patterns. O_2Si type. Quartz group. Also called: silica. Also called: low quartz. Silicon used as internal standard. PSC: hP9. To replace 5-490 and validated by calculated pattern. Plus 6 additional reflections to 0.9089.

dÅ	Int	hkℓ	dÅ	Int	hkℓ
4.257	22	100	1.1532	1	311
3.342	100	101	1.1405	<1	204
2.457	8	110	1.1143	<1	303
2.282	8	102	1.0813	2	312
2.237	4	111	1.0635	<1	400
2.127	6	200	1.0476	1	105
1.9792	4	201	1.0438	<1	401
1.8179	14	112	1.0347	<1	214
1.8021	<1	003	1.0150	1	223
1.6719	4	202	0.9898	1	402
1.6591	2	103	0.9873	1	313
1.6082	<1	210	0.9783	<1	304
1.5418	9	211	0.9762	1	320
1.4536	1	113	0.9636	<1	205
1.4189	<1	300			
1.3820	6	212			
1.3752	7	203			
1.3718	8	301			
1.2880	2	104			
1.2558	2	302			
1.2285	1	220			
1.1999	2	213			
1.1978	1	221			
1.1843	3	114			
1.1804	3	310			

Fig. 4. Example of an image in the PDF-2 (Powder Diffraction File).

sets (**Fig. 4**), each of which contains a list of *d/I* pairs, a chemical formula, a name, a unique identification (PDF) number, and a reference to the primary source. Supplemental data may be added, including Miller indices for all lines, unit cell and space group data, physical constants, and experimental details. Because the number of patterns is large, special ways of organizing the *d*'s and *I*'s into subfiles have been devised, such as minerals, zeolites, and pharmaceuticals. A smaller version of the database called PDF-1 is produced mainly for search programs on minicomputers with limited disk storage. As of 2001, about 160,000 patterns of the PDF-2 database require about 620 megabytes of storage, and PDF-1 about 39 megabytes.

Instrumentation. There are many types of powder diffractometer available ranging from simple laboratory instruments to versatile and complex instruments using a synchrotron source. Specialized instruments allow recording of diffraction patterns under nonambient conditions, including variable temperature, pressure, and atmosphere. Completely automated equipment for x-ray analysis is available. Most laboratory instruments consist of a high-voltage generator which provides stabilized voltage for the x-ray tube, so that the x-ray source intensity varies by less than 1%. A diffractometer goniometer is mounted on a table in front of the x-ray tube window. Electronic circuits use an x-ray detector to convert the diffracted x-ray photons to measurable voltage pulses, and to record the diffraction data.

The number of elements useful for x-ray tube targets is limited to a few, of which copper is the most commonly used. The copper K spectrum consists of a few lines superimposed on a broad continuum. These lines are the intense Cu $K\alpha_1/K\alpha_2$ doublet and the weaker Cu $K\beta_1/\beta_3$ doublet. In the α doublet, the $K\alpha_1$ line at 0.1541 nanometer is about twice as intense as the $K\alpha_2$ line at 0.1544 nm. The individual components of the $K\beta_1/\beta_3$ doublet are not usually resolved, and appear as a single line of average wavelength 0.1392 nm. The continuum begins at a wavelength which depends on the applied voltage, rises rapidly to a broad maximum, and then diminishes with increasing wavelength. The intensity of the continuum is mainly dependent on the x-ray tube voltage, but it also increases with atomic number of the target element. The entire spectrum is scattered and diffracted with varying efficiencies by the polycrystalline specimen. The interpretation of the diffraction pattern is simplified if the $K\beta$ line is eliminated. This has traditionally been accomplished by inserting a β-filter in the x-ray path. The filter consists of a thin nickel foil which almost completely absorbs the $K\beta$ radiation and transmits the $K\alpha$. More modern diffractometers employ a primary or diffracted beam monochromator, or a proportional Si(Li) detector to achieve the same result.

The two principal methods of recording diffraction patterns are by the use of film and with x-ray counter tubes. In the film method (**Fig. 5***a*), a small amount of the powder is glued to a thin glass fiber and rotated continuously in the center of a powder camera to give better randomness. The latter is essentially a light-tight cylindrical enclosure, usually about 4.5 in. (11 cm) in diameter and provided with collimators to define the x-ray beam. The strip of x-ray film is placed against the inside circumference of the cylinder concentric with the specimen axis. Two holes are punched in the film, one to admit the

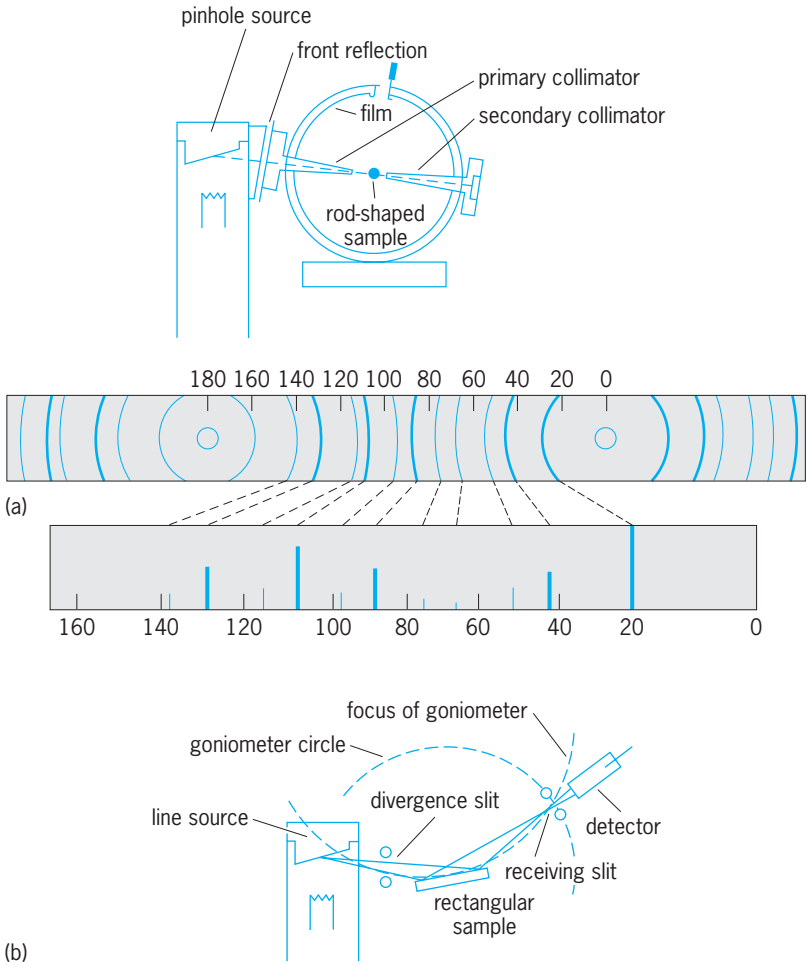


Fig. 5. Instrumentation for the measurement of x-ray powder patterns. (*a*) Debye-Scherrer camera. (*b*) Powder diffractometer.

collimator, and the other to allow the undiffracted beam to pass through. Exposures of 1–4 hours are required. Each properly oriented crystallite may produce a reflection that appears as a spot on the film. For each set of atomic planes there will be a large number of crystallites of various orientations, each producing its own spot, and all the reflected rays will form the surface of a cone, with its apex at the specimen and subtending the angle 4Θ on the film. The various sets of atomic planes thus produce a series of concentric cones, and these appear on the film as a series of arcs whose curvature depends on the reflection angle. After the film is developed and dried, it is laid flat and the linear distance is measured from the position on the film where the direct beam passed through to each of the arcs. Since, the wavelength λ and the camera diameter are known and since allowance may be made for film shrinkage in development, these linear distances may be converted into angular values, from which d may be calculated for each set of planes by substituting in the Bragg equation.

The more common diffractometer method employs a counter tube instead of film (Fig. 5*b*). The specimen is prepared with a flat surface which is exposed to the x-ray beam. The counter tube always points toward the specimen, and the reflection data are recorded line by line. The divergent primary x-ray beam converges, after reflection from the specimen, to a narrow slit in front of the counter tube. By rotating the counter tube at twice the angular speed of the specimen so that the specimen surface always makes an angle Θ with the primary beam when the counter tube is at 2Θ, sharp well-resolved lines are obtained at all reflection angles. This instrument makes possible the direct and accurate measurement of intensities and reflection angles. The peak-to-background ratios, resolution, and line shapes are far superior to those obtained by the cylindrical film camera. The operation is completely automatic.

Phase identification. Since every crystalline material gives, at least in principle, a unique x-ray diffraction pattern, study of diffraction patterns from unknown phases offers a powerful means of qualitative identification by comparing an x-ray pattern from the material to be analyzed with a file of single-phase reference patterns. Although the potential for qualitative phase identification was certainly recognized from the very early days of x-ray diffraction, the first attempts to list reference patterns were not published in detail until the late 1930s, along with means of archiving and retrieval of patterns. These methods still provide the basis of many search/match methods in use today. All databases require some type of index system to allow access to information contained within the database. For searching print databases, it is common practice to supply a combination index plus search manual for each of the main subsets of the PDF. A number of manual searching methods have been developed based on the three methods in common use: the alphabetical method, the Hanawalt method and the Fink method. The alphabetical index is designed to permit a rapid systematic search for all patterns with a specified chemical content. The Hanawalt method involves grouping the patterns in the PDF according to the *d*-value of the 100% intensity line. Each interval is sorted on the *d*-value of the second most intense line. Subsequent lines are listed in order of decreasing intensity. The Fink method indexes a pattern on its eight largest *d*-spacing lines, and eight separate entries are made using a cyclic permutation of *d*-values. An *Alphabetical Index* and a *Hanawalt Search Manual* are published annually, and a *Fink Method Search Manual* is published at irregular intervals. **Table 2** illustrates the various entry methods employed in the common indexes. The *Alphabetical Index* is a chemistry-based index using only elemental information. The *Hanawalt Index* is an intensity-driven index since it employs only the strongest lines for searching. The *Fink Index* is a *d*-spacing-driven index since it employs mainly the largest *d*-values. There have been numerous attempts to completely automate the search-match process resulting in several very successful commercial products, which permit routine external and internal standard calibration, precision alignment checks, and new levels of accuracy and precision in data collection and analysis.

Quality of experimental data. The ability to recognize a reference pattern in an unknown material

TABLE 2. Types of PDF data searching indexes

Index	Entry method	Search parameters
Alphabetic	Chemistry	Permuted elemental symbols
Hanawalt	*I*/*d*	Three strongest lines
Fink	*d*/*I*	First eight lines
EISI	Chemistry/*d*	Low/high *Z* elements; *d*-spacing
Boolean	Various	*d*-spacings, chemistry, strong lines, CODEN,* physical properties, functional groups, and so on

* Designation assigned to a periodical title by the Chemical Abstracts Service.

strongly depends on the quality of the *d*'s and *I*'s in both the reference material and the unknown sample. One of the principal problems in the identification of materials by comparison of an experimental pattern with reference patterns is the variability in the quality of the data. For Debye-Scherrer camera data, one should assume an error window of $\pm\Delta 2\Theta = 0.1^\circ$; for normal diffractometer data, one typically assumes a $\pm\Delta 2\Theta = 0.05^\circ$; and for internal standard corrected diffractometer or Guinier camera data, a $\Delta 2\Theta$ window as low as 0.01° may be assumed. Although the experimentally measured parameter is generally the 2Θ value, the search-match parameter is invariably the *d*-value. Unfortunately, the error relationship between 2Θ and *d* is nonlinear. The most useful lines for phase identification are the low-angle lines, and these are also the lines subject to the largest error in *d*. As the quality of both reference and experimental patterns improves, the problem of pattern recognition becomes easier. There has been a major effort to meet the ever-increasing demand for the higher-quality data needed because of improved instrumentation and better techniques. The introduction of the computer for data collection, treatment, and processing has improved the quality of measured *d*-spacings, leading to an ongoing need for improvement in the quality of reference patterns. For example, the modern automated powder diffractometer can produce *d*-spacing accuracies of about 1 part per 1000 for all but the larger *d*-values. This quality of data corresponds to an average angular error of 0.01–0.03° in 2Θ.

Ron Jenkins

Bibliography. D. L. Bish and J. E. Post, *Modern Powder Diffraction*, Mineralogical Society of America, 1989; B. D. Cullity, *Elements of X-ray Diffraction*, 2d ed., Addison-Wesley, 1978; R. Jenkins and R. L. Snyder, *Introduction to X-ray Powder Diffractometry*, Wiley, 1996; H. P. Klug and L. E. Alexander, *X-ray Diffraction Procedures: For Polycrystalline and Amorphous Materials*, 2d ed., 1974.

X-ray spectrometry

A rapid and economical technique for quantitative analysis of the elemental composition of specimens. It differs from x-ray diffraction, whose purpose is the identification of crystalline compounds. It differs from spectrometry in the visible region of the spectrum in that the x-ray photons have energies of thousands of electronvolts and come from tightly bound inner-shell electrons in the atoms, whereas visible photons come from the outer electrons and have energies of only a few electronvolts. *See* X-RAY DIFFRACTION.

In x-ray spectrometry the irradiation of a sample by high-energy electrons, protons, or photons ionizes some of the atoms, which then emit characteristic x-rays whose wavelength λ depends on the atomic number *Z* of the element ($\lambda \propto 1/Z^2$), and whose intensity is related to the concentration of that element. Generally speaking, the characteristic x-ray lines are independent of the physical state (solid or liquid) and of the type of compound (valence) in which an element is present, because the x-ray emission comes from inner, well-shielded electrons in the atom. **Figure 1** illustrates the removal of one of the innermost, K-shell, electrons by a high-energy photon. The photon energy must be greater than the binding energy of the electron; the difference in energy appears as the kinetic energy of the ejected electron. The K-ionized atom is unstable, and one of the L- or M-shell electrons drops into the K-shell vacancy within 10^{-14} to 10^{-6} s. As this transition occurs, a characteristic x-ray photon is emitted with an energy equal to the difference in energy between the K and the L (or M) shell, or an additional electron, called an Auger electron, is ejected from the atom. Either the x-rays or the Auger electrons may be used for analysis, but in this article the discussion is concerned exclusively with the x-rays. *See* AUGER EFFECT.

Figure 2 shows some of the allowed transitions and the naming of the lines. There is a selection rule in atomic physics which says that only certain ones of the outer electrons are allowed to fill a vacancy in an inner shell. The rule can only be stated in terms of quantum mechanics: the transition must be from one shell to another, and the second (orbital) quantum number must change by ± 1, that is, $p \rightleftarrows s$, $d \rightleftarrows p$, and so on. **Figure 3** shows the λ versus *Z* relationship for the strongest *K*- and *L*-series lines. Figure 2 shows that some of the transitions may come from the valence shell, in which case there will be slight alteration of wavelength or line shape of characteristic

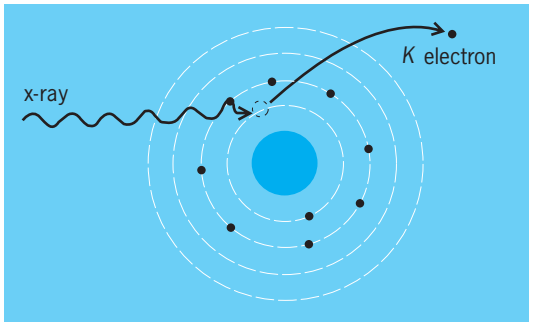


Fig. 1. Removal of a *K* electron from an atom by a primary x-ray photon. (*After L. S. Birks, X-ray Spectrochemical Analysis, 2d ed., Wiley-Interscience, 1969*)

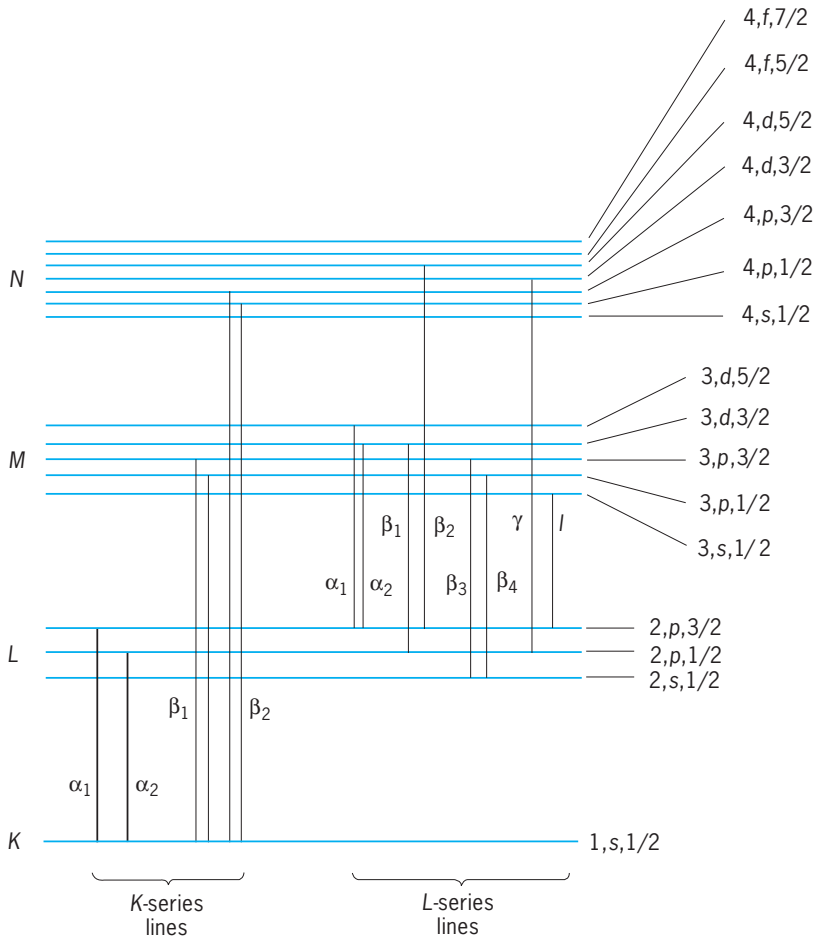


Fig. 2. Partial energy-level diagram showing the transitions leading to the *K* and *L* series lines. (*After L. S. Birks, X-ray Spectrochemical Analysis, 2d ed., Wiley-Interscience, 1969*)

elements of atomic number Z greater than 20, a pentaerythritol crystal for Z between 13 and 20, and a potassium acid phthalate crystal for Z between 8 and 13. X-ray fluorescence analysis is not generally suitable below about $Z = 9$, but with special instruments and techniques it can be extended down to about $Z = 5$.

The detectors used for wavelength dispersion are either gas proportional counters or scintillation counters. Both of these count individual photons, but the gas counters are most suitable for wavelengths longer than about 0.2 nanometer, while the scintillation counters are most suitable for shorter wavelengths. The amplitude of the output pulse for each photon is proportional to the energy of the x-ray photon it represents. However, the statistical variation in the amplitude for each specific photon energy means that characteristic lines from neighboring elements are not resolved by either gas proportional

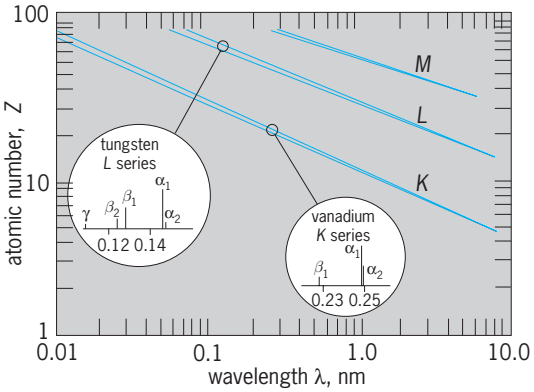


Fig. 3. Atomic number Z versus wavelength λ of the characteristic lines ($\lambda \propto 1/Z^2$). (*After L. S. Birks, X-ray Spectrochemical Analysis, 2d ed., Wiley-Interscience, 1969*)

lines with valence. This alteration can be measured for multivalent elements such as sulfur by using spectrometers designed for high resolution (**Fig. 4**).

Spectrum analysis. Photon generation of characteristic spectra is the most common and is called x-ray fluorescence. It is carried out with an x-ray tube as the source of primary radiation. There are two ways of analyzing the spectra: wavelength dispersion and energy dispersion.

Wavelength dispersion. This is shown in **Fig. 5*a***. The characteristic emission from the sample is usually excited by a chromium or tungsten target x-ray tube which is operated at 2–3 kW. The emitted radiation is limited to a parallel beam by the blade collimator and is diffracted, one wavelength at a time, by an analyzer crystal. Bragg's law ($n\lambda = 2d \sin \theta$) relates the diffraction angle θ to the wavelength λ for crystal planes with an interatomic-spacing distance d. The term n is the order of the diffraction, 1, 2, 3, etc., as it is in classical optics. As shown in Fig. 3, the characteristic wavelength decreases as the atomic number increases. The $2d$ spacing of the analyzing crystal must be greater than the wavelength being diffracted, but if it is too much greater the spectral lines will be crowded toward small θ. It has become the practice to use a crystal of lithium fluoride for

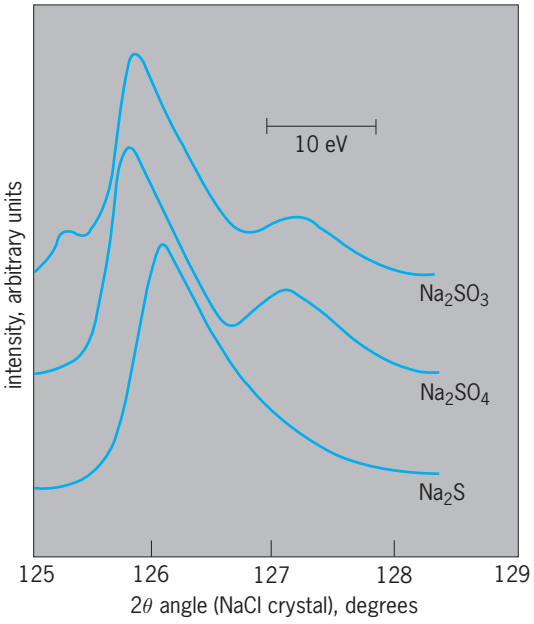


Fig. 4. Variations of the sulfur $K\beta$ line with valence state of sulfur. (*After L. S. Birks and J. V. Gilfrich, X-ray fluorescence analysis of the concentration and valence state of sulfur in pollution samples, Spectrochim. Acta, 33B, no. 7:305, 1978*)

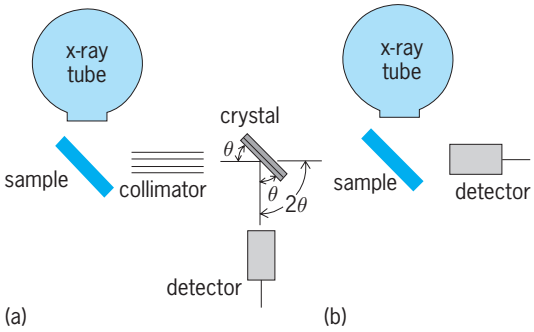


Fig. 5. Wavelength and energy dispersion methods. (*a*) Wavelength dispersion with a crystal spectrometer. (*b*) Energy dispersion with a solid-state detector.

or scintillation counters. **Figure 6** shows the resolution for the gas proportional counter alone and with a crystal spectrometer, as well as for the silicon solid-state detector used for energy dispersion analysis, which is discussed below; resolution with a scintillation counter is almost a factor-of-3 worse than a gas proportional counter. In wavelength dispersion it is the resolution of the crystal spectrometer, however, which determines the separation between neighboring wavelengths, and the crystal resolution is better than any of the detectors.

Energy dispersion. In this method all of the radiation emitted by the sample enters an energy-sensitive detector, usually a silicon solid-state detector (Fig. 5*b*). Such detectors are operated at liquid-nitrogen temperature to reduce electronic noise and allow an energy resolution of about 150 eV (Fig. 6). This resolution is adequate to distinguish the $K\alpha$ and $K\beta$ lines of a single element, but not adequate to separate the $K\beta$ line of one element from the $K\alpha$ line of the next higher atomic number element (for example, Cr$K\beta$ from Mn$K\alpha$). **Figure 7** compares the wavelength and energy-dispersion spectra from a sample containing a range of atomic number elements. Resolution is better with the crystal spectrometer for most of the spectral range of interest.

In spite of its relatively poor resolution, energy dispersion has become widely accepted because of two advantages it has over wavelength dispersion. First, all the characteristic lines are recorded simultaneously, which makes the method faster than a scanning crystal spectrometer (but not as fast as the multiple crystal-spectrometer instruments). Second, the solid angle of radiation accepted by the detector is 10–100 times greater than the solid angle accepted by the crystal; this allows the primary-source power to be reduced proportionally. With energy dispersion, it is feasible to use electron excitation at beam currents below 10^{-8} ampere in the scanning electron microscope, compared to beam currents of about 10^{-6} A required in electron probes used with crystal spectrometers. Likewise, with energy dispersion, it is feasible to use proton or alpha-particle beams from Van de Graaff or cyclotron accelerators. For proton beams, the proton energy must be 1–5 MeV, compared to electron energies of 10–50 keV for the same x-ray yield; for alpha particles, the energy should be even higher, 10–50 MeV. Compared to photon excitation, positive-ion excitation results in lower background intensity and improves the limit of detection by as much as a factor of 10. On the other hand, direct electron excitation results in much higher background intensity because it is generated in the sample rather than merely scattered by the sample; this degrades the limit of detection by a factor of 10–100.

With energy dispersion, it is even feasible to use some of the radioactive isotope sources as the primary radiation to excite the characteristic x-ray spectra in the sample. There are isotopes such as tritium which undergo beta decay and, when used in a metal matrix, produce continuum and characteristic matrix x-rays; other isotopes such as americium-241 undergo alpha decay and produce gamma rays or x-rays. The advantage of isotope sources is their compact size and the elimination of power supplies and so forth. However, advances in low-power air-cooled x-ray tubes (10–50 W) and compact power supplies have largely eliminated the need for isotope sources.

Types of samples and data interpretation. Two general classes of samples are analyzed routinely by x-ray spectrometry: thin and bulk samples.

Thin samples. Samples smaller than several milligrams per square centimeter such as air-pollution particles collected on a filter, constitute such a thin layer that the primary radiation can penetrate easily and excite emission from each element in proportion to the mass per unit area of that element. Quantitative calibration for each element is accomplished experimentally by determining the sensitivity S in photons per second per microgram per square centimeter for that element. Accuracy (or more properly, precision) depends on the number N of photons counted; the

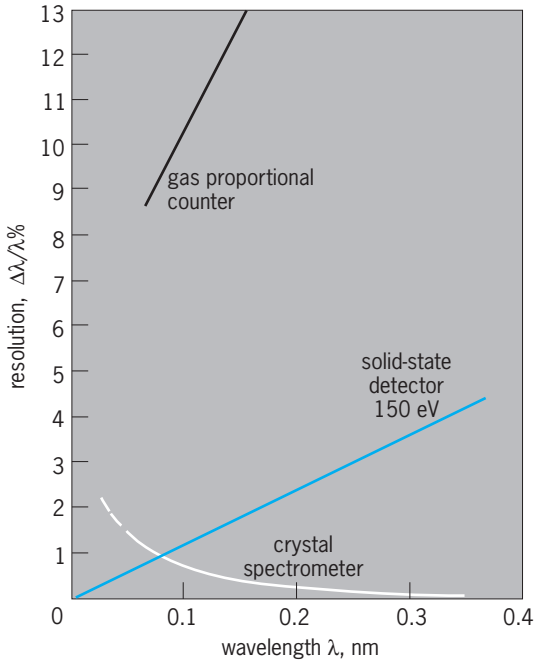


Fig. 6. Resolution of a solid-state detector and a gas proportional counter alone or in conjunction with a crystal spectrometer.

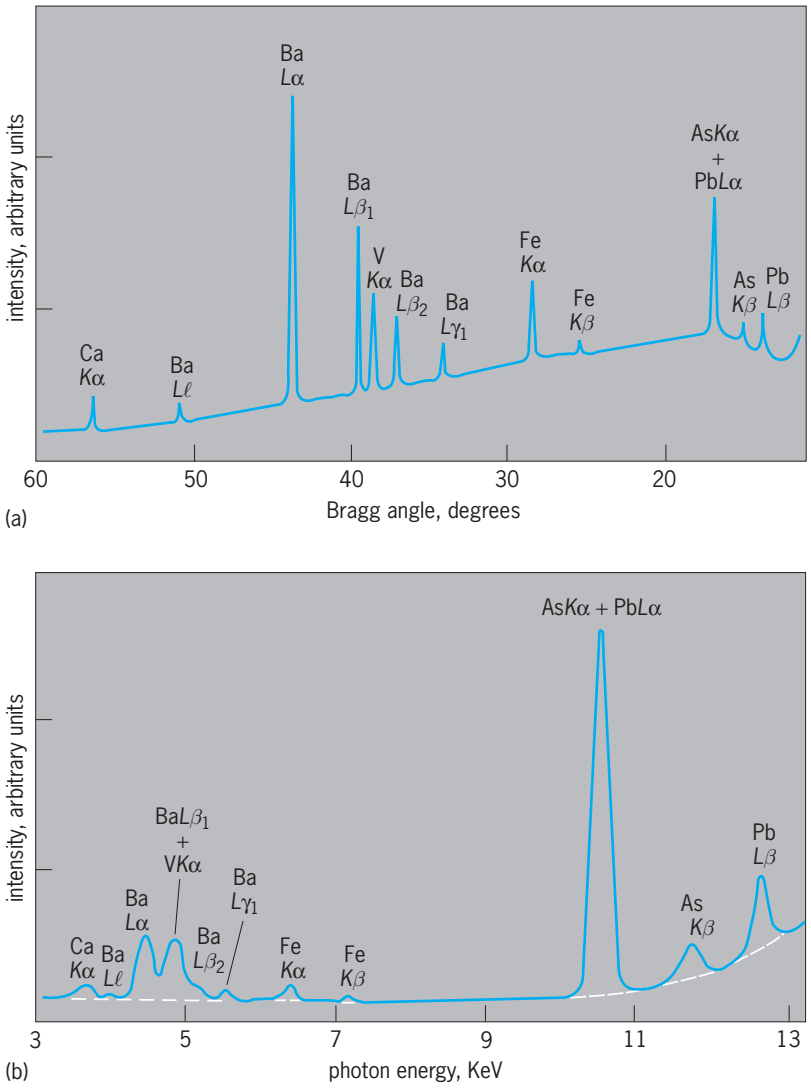


Fig. 7. X-ray spectra of the same sample as measured by (*a*) wavelength dispersion and (*b*) energy dispersion. (*After L. S. Birks, Pinpointing airborne pollutants, Environ. Sci. Tech., 12:150, 1978*)

expected standard deviation σ of a measurement is approximated by $\sigma = \sqrt{N}$. Limit of detection C_L depends not only on sensitivity but on the background intensity as well. If the measured intensity of the background is N_B and that at the line peak is N_p, the limit of detection is defined as the amount of material C_L which gives a signal above background of $3\sigma_B$, that is, $N_p - N_B = 3\sqrt{N_B}$ and $C_L = 3\sqrt{N_B}/(S \times t)$, where t is the counting interval in seconds. Thus it is important to minimize the background intensity N_B by careful instrumental design to eliminate scattering from material other than the sample, and also to minimize the mass per square centimeter of the substrate on which the sample is mounted. With photon excitation and optimized design, the limit of detection varies from about 1 nanogram/cm^2 for elements around atomic number 20 to 10–50 ng/cm^2 for the extremes of high or low Z. The limit is approximately the same for either wavelength or energy dispersion when photon excitation is used. As was stated in the previous portion of this subject area, the limit of detection can be improved by about a factor of 10 with proton excitation, but the analysis is then generally limited to energy dispersion.

Bulk samples. These consist of solids, powder, or liquids and present quite a different problem from thin samples. X-ray absorption limits the penetration of the primary radiation through the sample matrix and the depth from which characteristic radiation may emerge. In addition, the characteristic radiation from some of the elements in the specimen may excite the characteristic radiation of other elements by secondary fluorescence. Matrix absorption and secondary fluorescence depend on the sample composition and determine the x-ray intensity versus composition relationship for each element. Thus, instead of a single, linear calibration as was described for thin samples, the calibration is generally nonlinear, and a family of curves is needed for each element as the matrix composition changes.

If I_{ai} is the intensity from a pure sample of element i, and I_{bi} is the intensity from an unknown concentration of element i in a matrix of other elements, then the most useful parameter is the relative x-ray intensity $R_i = I_{bi}/I_{ai}$. To a first approximation, R_i can be expressed in terms of the concentration C_i and the absorption term μ which incorporates both the absorption of the incident primary radiation and the emerging characteristic radiation, as shown in Eq. (1), where A is a constant which depends on a

$$R_i = \frac{AC_i}{\mu} \tag{1}$$

number of instrumental parameters.

In x-ray spectrometry the analyst can measure R_i but cannot determine C_i directly, because μ depends on C_i and the concentration of each other element C_j as well. However, it is possible to write an equation containing R_i and C_i in terms of individual absorption and secondary-fluorescence effects of each element on the intensity from each other element. The expression is Eq. (2), where Σ is the sum of a

$$\frac{C_i}{R_i} = 1 + \Sigma\alpha_{ij}C_j \tag{2}$$

number of terms, one for each other element in the specimen. The coefficient α_{ij} means the effect on element i by the presence of element j; it is often referred to as an influence coefficient and may be determined experimentally by measuring mixtures of elements i and j. The terms may also be calculated from some of the fundamental properties of atoms and radiation.

There are many variations of the mathematical expressions for determining concentration from x-ray intensity; most of them require computers to evaluate. Whatever the method of data interpretation, a few generalizations may be made about the analysis of bulk specimens. The accuracy (precision) is about 1–2% of the amount of an element present for major constituents, but degrades to 5–10% of the amount present for concentrations to 10–100 ppm. The limit of detection is generally about 1 ppm for

middle-range atomic number elements, but varies from less than 0.1 ppm for metals in biological tissue to 10 ppm or more for low-*Z* elements such as carbon in a middle-*Z* matrix such as steel.

Improvements and limitations. X-ray spectrometry generally does not require any separation of elements before measuring, because the x-ray lines are easily resolved. However, preconcentration methods are sometimes useful as a means for improving the limit of detection. An example is the precipitation or ion-exchange collection of soluble elements in water. Likewise, dilution is sometimes useful to reduce matrix variability or inhomogeneity. An example is the solution of mineral samples in borax.

One limitation of x-ray spectrometry is the progressive difficulty of measurement below atomic number 11. There are several reasons for this, including the strong absorption of such long-wavelength radiation by the x-ray tube and detector windows and also the reduced intensity due to a lower number of x-ray photons emitted per atom ionized (the fluorescent yield factor). Although the elements from boron (5) to fluorine (9) may be measured with specially designed equipment, such measurement cannot be considered routine. In practice, photoelectron spectroscopy and Auger electron spectroscopy are favored for the lower-atomic-number elements, but electron methods may also require special instrumentation and techniques. *See* ELECTRON SPECTROSCOPY; SPECTROSCOPY.

L. S. Birks

Bibliography. B. K. Agarwal, *X-ray Spectroscopy*, 2d ed., 1991; E. P. Bertin, *Principles and Practices of X-ray Spectrometric Analysis,* 2d ed., 1975; K. Debertin and R. G. Helmer, *Gamma- and X-ray Spectrometry with Semiconductor Detectors*, 2005; R. E. Van Grieken and A. A. Markowicz (eds.), *Handbook of X-ray Spectrometry Revised and Expanded* (*Practical Spectroscopy, V. 29*), 2d ed., 2001; A. Meisel, *X-ray Spectra and Chemical Binding,* 1989; K. Tsuji, J. Injuk, and R. Van Grieken (eds.), *X-ray Spectrometry: Recent Technological Advances*, 2004.

X-ray telescope

An instrument designed to collect and detect x-rays emitted from a source outside the Earth's atmosphere and to resolve the x-rays into an image. Absorption by the atmosphere requires that x-ray telescopes be carried to high altitudes. Balloons are used for detection systems designed for higher-energy (hard) x-ray observations, whereas rockets and satellites are required for softer x-ray detectors. *See* X-RAY ASTRONOMY; X-RAYS.

Image formation. An image-forming telescope lens for x-ray wavelengths can be based either on the phenomenon of total external reflection at a surface where the index of refraction changes (grazing-incidence telescope) or on the principles of constructive interference (multilayer telescope).

Grazing-incidence telescope. In the case of x-rays, the index of refraction in matter is slightly less than unity. By applications of Snell's law, the condition for total external reflection is that the radiation be incident at small grazing angles, less than a critical angle of about 1°, to the reflecting surface. The value of the critical angle depends on the wavelength of the radiation and the material used. As the wavelength decreases (higher energy), the grazing angle required is smaller; as the atomic number (*Z*) of the material increases, the grazing angle for a given wavelength increases. The detailed reflectivity as a function of energy is complicated by x-ray absorption edges in the material, as well as the density of the material and its surface properties. *See* REFLECTION OF ELECTROMAGNETIC RADIATION; REFRACTION OF WAVES.

Based on these properties, x-ray mirrors have been constructed which focus an image in two dimensions. Various configurations of surfaces are possible. In the Wolter type I x-ray telescope (**Fig. 1**), which has been built and used successfully for x-ray astronomy on the *Einstein* and *ROSAT* observatories, two surfaces of revolution are used to produce a high-quality image. These mirrors are manufactured from fused quartz, which is coated with nickel after being shaped and polished. The surfaces are extremely smooth in order to focus well, and the shape of the mirror must also be within narrow tolerances. The *Einstein Observatory* mirrors were the first ever built for nonsolar x-ray imaging. They were four nested pairs ranging about 2–1 ft (0.6–0.3 m) in diameter and 4 ft (1.2 m) in length, and the total area of the polished surface was about the same as that of the 200-in. (5-m) telescope at Palomar Mountain; yet the effective area of the telescope was only about 150 in.^{2} (1000 cm^2), because of the grazing-angle geometry.

Both the *Einstein* and *ROSAT* observatories were limited by their optical designs to very low energies. For *Einstein* the telescope cut off at energies above 4.5 keV; that is, x-rays of higher energy were not reflected because their grazing angle (set by the shape of the telescope and its focal length) was too large. For *ROSAT* the high-energy cutoff occurred at 2.5 keV. Subsequent telescopes had higher cutoffs: 12 keV for ASCA, 10 keV for *BeppoSAX* and *Chandra*, and 15 keV for *XMM-Newton*.

Multilayer telescope. A second type of x-ray telescope is based on the principles of constructive

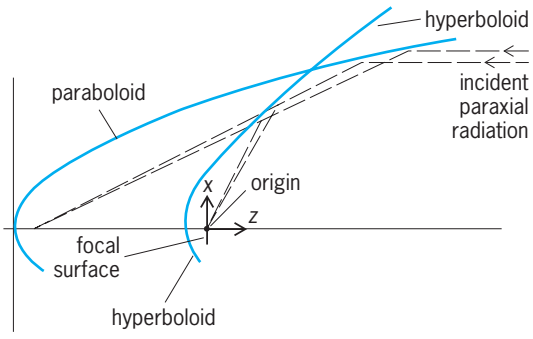


Fig. 1. Wolter type I x-ray telescope. Two surfaces of revolution (paraboloid and hyperboloid) reflect the incident x-rays to a common focus. The horizontal scale is greatly compressed. The *y* axis is perpendicular to the plane of the figure.

interference in extremely thin layers of material deposited on a mirror surface. Unlike the grazing-incidence telescope, multilayer telescopes do not require the x-rays to strike at shallow angles in order to be reflected. Instead these mirrors are similar to normal optical telescope mirrors where the incoming radiation strikes the mirror at nearly normal incidence to be reflected and focused. *See* INTERFERENCE OF WAVES.

The multilayers are coatings of specially selected materials that have crystal structures of regularly spaced atoms. They are evaporated onto a mirror surface that has been highly polished so that it is the correct shape to focus x-rays and is so smooth that the average bumpiness on the mirror surface is comparable to the wavelength of the x-rays being reflected. This is the same quality of surface that is required for grazing-incidence mirrors, but is somewhat easier to achieve when the surface being polished is nearly flat. Several layers of coatings are used in order to achieve high reflectivity, and each layer must be of a precise thickness and composition to work properly. These coatings are deposited in vacuum to achieve the desired high degree of cleanliness and purity of the multilayers if they are to work efficiently. *See* ARTIFICIALLY LAYERED STRUCTURES.

The advantage of a multilayer mirror is that all of the mirror area is used in collecting the x-ray radiation, whereas the grazing-incidence telescopes have only a small projected area of the actual mirror surface collecting radiation. There is an offsetting disadvantage, as the multilayer mirror reflects x-rays only within a very narrow range of energy while the grazing-incidence mirror reflects over a broad range of energies. The effect is similar to using a narrow-band filter with an optical telescope, and in some cases this can be very useful.

The first astronomical use of multilayer, normal-incidence x-ray telescopes was in photographing the Sun. Images of the solar surface of very high angular resolution were made in a selected wavelength interval near 6.7 nanometers. They showed that the filamentary structure of solar flares extends down to as small a size as can be resolved in the pictures, a result that has direct consequences for the theoretical work being done on understanding these flares. The Sun is an excellent object for observation with multilayer telescopes because it is so bright (being nearby) that there is enough flux for detection over even a narrow energy band.

Image detection. The telescope mirrors focus x-rays, producing an image in two dimensions in the same manner as the lenses of optical telescopes result in images of the sky. Suitable devices are required to detect and record these images, completing the functional requirements of an astronomical telescope. Various types of x-ray detectors have been developed for this purpose. These are position-sensitive devices which in effect are electronic cameras suitable for x-ray wavelengths.

Microchannel plates. The high angular resolution of the grazing-incidence telescope requires a camera that has correspondingly good spatial resolution. One type of detector uses microchannel plates (MCP; **Fig. 2**) and yields about 20 micrometers resolution for x-rays in the soft energy band about 100–10,000 eV). The microchannel plate is an array of small hollow tubes or channels (about 10–15 μm in diameter) which are processed to have high secondary electron yield from their inner walls. A single x-ray striking the surface of a channel may produce a free electron. The electric field produced by placing a high voltage across the microchannel plate accelerates this electron, which collides with the wall of the tube to produce more electrons. This results in a cascade of electrons through the channel, multiplying in number until a sufficient signal is produced to be recorded electronically, giving the location of the event. *See* IMAGE TUBE (ASTRONOMY); PHOTOMULTIPLIER.

Charge-coupled devices. Another type of detector developed for x-ray imaging applications is the charge-coupled device (CCD). This is a solid-state detector that consists of microscopic silicon picture elements (pixels) in which electronic charges produced by the passage of an x-ray photon are collected. The charge-coupled device is periodically read out by shifting the collected charges from one pixel to the next, as in a bucket brigade. The charge is detected in a sensitive amplifier, and the corresponding position is determined by keeping count of the number of charge transfers that were made. Typical charge-coupled devices have pixels that are 15–25 μm on a side, and there are up to 4096 × 4096 pixels on one such device. The charge-coupled device not only records the position of an event but can also yield information on the energy of the photon. Thus this type of detector is useful as an imaging spectrometer, providing a spectrally resolved image of an x-ray source

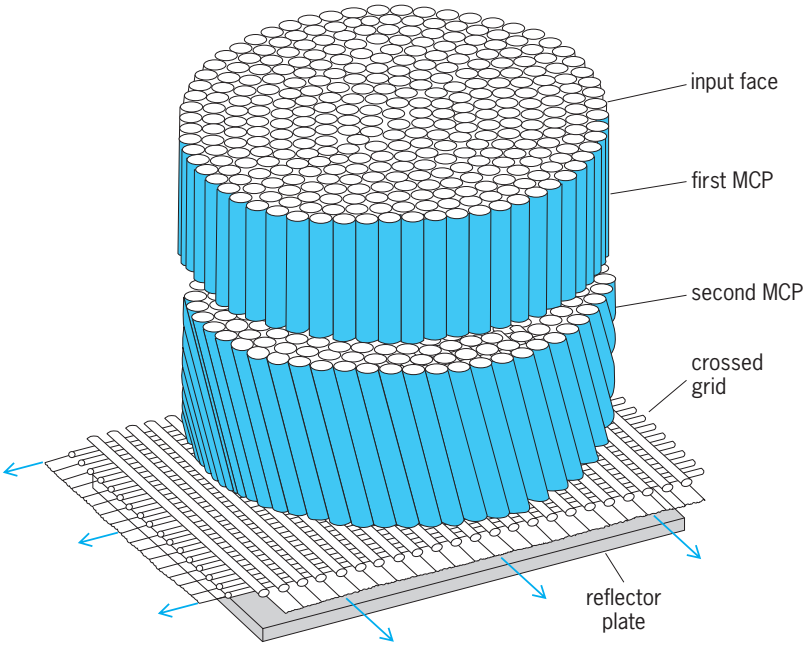


Fig. 2. Blow-up (not to scale) of the major components of a high-resolution imaging detector for an x-ray telescope. Microchannel plates (MCP) detect incident radiation and amplify the signal for position determination by the grid of crossed wires below.

which can be analyzed for the energy distribution of x-rays in different parts of the image. Such devices are needed to study supernova remnants and other extended sources in order to understand the conditions within such objects and how they vary. *See* CHARGE-COUPLED DEVICES.

Position-sensitive proportional counters. Yet another type of detector used with x-ray telescopes is a gas-filled counter in which x-rays are photoelectrically absorbed, yielding an electron which is detected by the ionization it produces in the gas. By operation of the counter in the proportional mode and use of planes of wires to localize the electrical signals, the position and amplitude of each event can be recorded. These detectors generally have lower spatial resolution (about 200 μm) than do microchannel plate or charge-coupled-device detectors, but they can be made larger in size and provide better energy resolution than microchannel plate detectors. Position-sensitive proportional counters were used in the *Einstein Observatory* and in *ROSAT*. *See* IONIZATION CHAMBER.

Cryogenic detectors. Another class of detectors consists of very low temperature devices that detect the heat deposited in an absorber when an x-ray is stopped. These devices promise improved energy resolution and efficiency over a broad range of energies. They all work at temperatures close to absolute zero, so that keeping the detector cold is a major problem. *See* TELESCOPE.

Stephen S. Murray

Bibliography. G. Burbidge and A. Hewitt, *Telescopes for the '80s*, 1981; G. W. Fraser, *X-ray Detectors in Astronomy*, 1989; L. Golub (ed.), *X-ray Instrumentation in Astronomy II*, 1989; R. B. Hoover and A. B. Walker (eds.), *X-ray Optics, Instruments, and Missions II*, 1999; P. Joss (ed.), *High-Energy Astrophysics in the Twenty-First Century*, 1990; O. H. Siegmund and K. A. Flanagan (eds.), *EUV, X-ray, and Gamma-ray Instrumentation for Astronomy*, vol. 10, 1999.

X-ray tube

An electronic device used for the generation of x-rays. X-rays are produced when electrons that have been accelerated by a large potential are stopped in a target, which is usually made of metal. The x-rays radiate in all directions from the spot on the target where the electrons hit. X-rays are produced by two mechanisms in a metal target: bremsstrahlung (or braking radiation in German), which produces a continuous energy spectrum of x-rays; and characteristic x-rays, which are produced by the creation and subsequent filling of a vacancy in an inner shell of the target atoms (**Fig. 1**). *See* BREMSSTRAHLUNG; X-RAYS.

Gas x-ray tubes. In gas tubes, electrons are freed from a cold cathode by positive ion bombardment. For the existence of the positive ions a certain gas pressure is required without which the tube will allow no current to pass. In the earliest gas tube the electrons liberated by positive ion bombardment

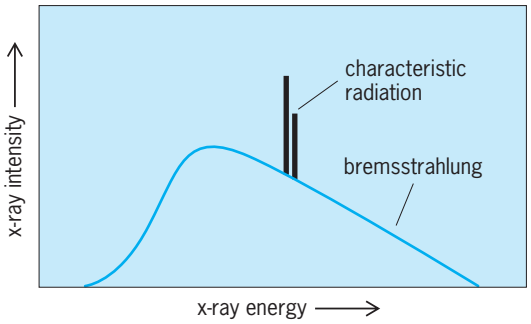


Fig. 1. X-ray spectrum from an x-ray tube.

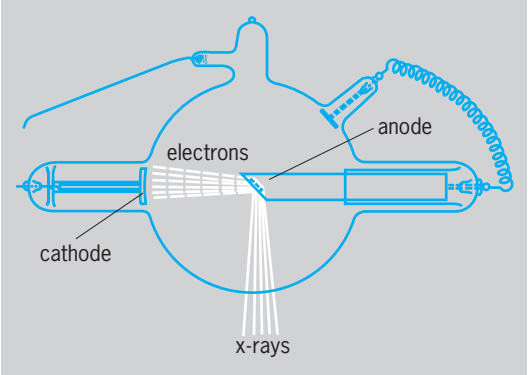


Fig. 2. Commercial model of gas x-ray tube.

from a flat aluminum cathode were emitted in a direction perpendicular to the cathode surface. The electrons traveled in straight lines until they impinged upon the glass end wall of the tube, where x-rays were generated.

Many designs of gas tubes have been built for useful application, particularly in the medical field. Metals, such as platinum and tungsten, have been placed in the path of the electron beam to replace glass as the target. Concave metal cathodes are used to focus the electrons on a small area of the metal target and increase the sharpness of the resulting shadows on the fluorescent screen or the photographic film. **Figure 2** shows one form of commercial gas x-ray tube.

The size of the bulb in commercial tubes was increased as the power input was increased, to reduce the pressure change in operation and to reduce the local heating of the glass envelope resulting from bombardment of electrons reflected from the focal spot. The aluminum cathodes were made heavier to withstand the increased positive-ion bombardment. The thin metal targets were also replaced by a heavier mass of metal. Targets now consist of two main parts: a refractory metal face, such as platinum, to take the direct impact of the electron beam, and a heavy backplate of a good heat-conducting material, such as copper, to conduct the heat away from the focal spot and store it temporarily.

Early in the development of x-ray tubes metals of high atomic weight were known to be the most efficient x-ray generators. W. C. Roentgen had used both aluminum and platinum as targets. W. D. Coolidge

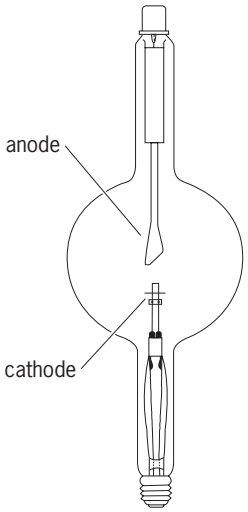


Fig. 3. Early commercial model of single-section, hot-cathode, high-vacuum x-ray tube.

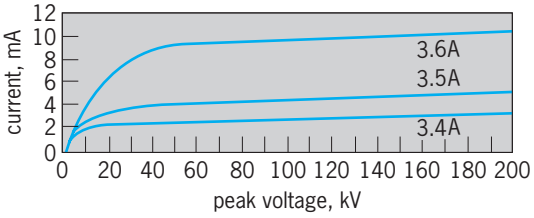


Fig. 4. Curves showing relation of current to voltage in hot-cathode, high-vacuum x-ray tube. Filament currents of the three curves are in amperes.

and others later used thorium and uranium with increased efficiency. The principal properties desired for a suitable target material are (1) high atomic number to give best x-ray production efficiency, (2) high melting point and high thermal conductivity to permit maximum electron power to be put into the focal spot, and (3) low vapor pressure to reduce the rate of evaporation of the metal on the walls of the glass envelope. Ductile tungsten as developed by Coolidge was found to combine these desired properties of an x-ray target to the greatest degree.

High-vacuum x-ray tubes. The operational difficulties and erratic behavior of gas x-ray tubes are inherently associated with the gas itself and the positive ion bombardment that takes place during operation. The high-vacuum x-ray tube eliminates these difficulties by using other means of emitting electrons from the cathode.

Coolidge made a high-vacuum x-ray tube with a hot tungsten-filament cathode and a solid tungsten target. This hot-cathode high-vacuum type of x-ray tube permitted stable and reproducible operation with relatively high voltages and large masses of metals. The vacuum was so good that positive ions did not play either an essential or a harmful role in the tube operation. The earliest commercial form of the hot-cathode high-vacuum type of x-ray tube (**Fig. 3**) utilizes a solid tungsten target. The independent relation of the current to the impressed voltage in a tube of this type is shown in **Fig. 4**. The different curves are for different filament temperatures and show that, over the operating range of x-ray voltages, the discharge current is practically independent of voltage. One typical form of a modern commercial hot-cathode high-vacuum x-ray tube (**Fig. 5**) is built with a liquid-cooled, copper-backed tungsten target, which operates over a wide range of energy ratings and is capable of rectifying its own current.

The inherent advantages of the hot-cathode high-vacuum tube over the gas tube are (1) flexibility—the voltage and current may be varied independently; (2) stability—this permits more accurate reproducibility of results; (3) small size; (4) operation—it can be operated directly from a transformer, making possible a very simple unit; and (5) long life.

High-voltage x-ray tubes. The operating voltage of a single-section hot-cathode x-ray tube is limited by the field current emitted from the cathode by the high electrostatic field. Field emission may cause erratic and uncontrollable operating behavior of an x-ray tube. Single-section x-ray tubes must preferably be operated at voltages below which field currents cannot be produced. *See* FIELD EMISSION.

There are, however, some highly specialized applications of x-ray tubes for which the tube is designed specifically to produce and use field emission currents to generate x-rays. To build x-ray tubes that operate stably and continuously at very high voltages, the multisection principle is usually employed. These tubes are made with many intermediate sections between the cathode and anode sections. The voltage applied across each section of this multisection tube is always less than that at which field currents will be produced. The electrons emitted from the cathode are accelerated by the voltage applied across each intermediate section. The sum of all of these sectional voltages determines the voltage rating of the multisection x-ray tube. By this procedure, x-ray tubes can be built to generate x-rays at several millionelectronvolts energy. **Figure 6** is a commercial 2-MeV x-ray tube built according to this principle.

The toroidal electromagnetic type of hot-cathode high-vacuum x-ray tube, first successfully built by D. W. Kerst, permits the generation of x-rays at energy levels of many megaelectronvolts without encountering the high-voltage insulation problem that

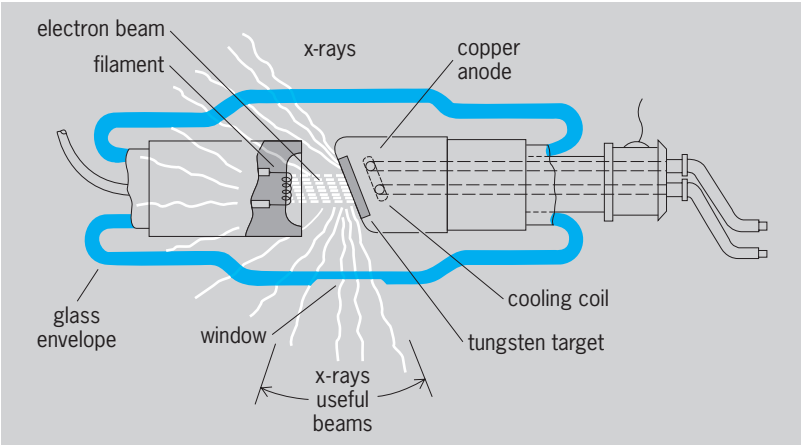


Fig. 5. Single-section, hot-cathode, high-vacuum x-ray tube.

is present in more conventional types of x-ray tubes. This x-ray tube consists of a toroidal high-vacuum envelope housing the hot-cathode and the x-ray target. This tube is designed to guide and accelerate electrons in a circular orbit in a machine called the betatron. In this toroidal-shaped betatron tube (**Fig. 7**) the electrons are injected and accelerated to an energy of many millions of electronvolts before they strike the target to produce x-rays. The average electron beam current is very small, but x-ray tubes of this type have been built to generate x-rays at energies ranging from 1 to over 100 MeV.

A synchrotron is a particle accelerator which can accelerate and store electrons at very high energies. X-rays are emitted when the direction of motion of relativistic stored electrons is changed, as by bending in a magnetic field. Synchrotron storage rings produce exceptionally bright beams of x-rays over a wide spectrum. *See* PARTICLE ACCELERATOR; SYNCHROTRON RADIATION.

The multisection linear accelerator type of x-ray tube is another hot-cathode high-vacuum device that is capable of generating intense x-rays over a wide range of energy levels from 1 to many megaelectronvolts. It consists (**Fig. 8**) of an electron gun and a copper pipe waveguide in which the electrons are accelerated by the axial electric field produced by a high-frequency oscillator. At the output end of the waveguide, the electrons strike a target and generate x-rays. X-ray tubes of this type have been built for operation at 5–15 MeV and higher, and with average currents of the order of 25 microamperes or more.

Applications. Many special forms of x-ray tubes, following the basic patterns already outlined, have been built for application in medicine, industry, and fundamental science. They vary from a few centimeters to many meters in length. X-ray tubes are used as an aid in medical diagnosis and in therapeutic treatment. They are used in the nondestructive testing of materials throughout industry. In both fundamental and applied science they are widely used in crystal-diffraction work, chemical analysis by x-ray spectra and absorption, and research on atomic structure. *See* MEDICAL IMAGING; NONDESTRUCTIVE EVALUATION; RADIOGRAPHY; RADIOLOGY; X-RAY CRYSTALLOGRAPHY; X-RAY DIFFRACTION; X-RAY FLUORESCENCE ANALYSIS; X-RAY POWDER METHODS.

To provide maximum use in these wide fields of application, particular design features are built into the tubes to meet the special requirements. Some of these special requirements are low or high currents, low- or high-voltage x-rays, large or small focal spot, high x-ray beam intensity for short or long intervals, and choice of x-ray target material to generate a particular quality of x-ray spectra. X-ray and electrical protection must be ensured.

The overall capacity of an x-ray tube is determined principally by the target material, the area of the focal spot, the duration of the energy applied, and the temperature of the target during the time the electron energy is applied. The modern x-ray tube is a precision tool of great stability and flexibility, capable of controlled operation with currents and voltages of any desired magnitude, and generating x-rays over a wide range of energy and intensity. Ernest E. Charlton

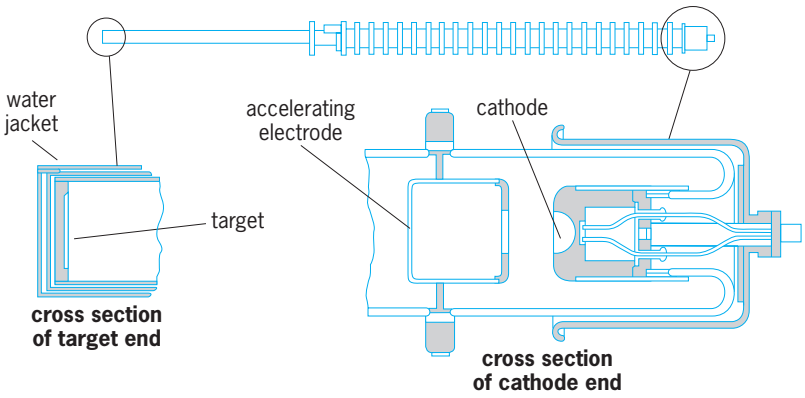


Fig. 6. A 2-MeV multisection, hot-cathode, high-vacuum x-ray tube.

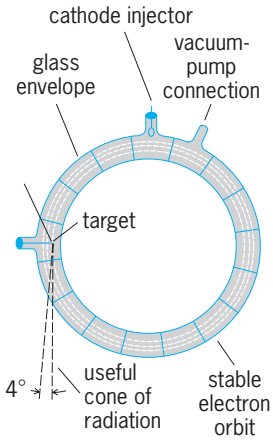


Fig. 7. Betatron "doughnut" high-vacuum x-ray tube. (*General Electric Co.*)

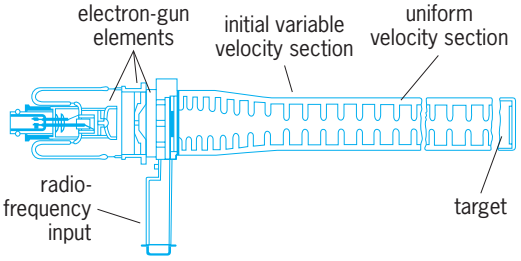


Fig. 8. Multisection linear accelerator type of hot-cathode, high-vacuum x-ray tube.

Bibliography. M. Y. Chen, T. L. Pope, and D. J. Ott, *Basic Radiology*, 2d ed., McGraw-Hill, 2004; R. Halmshaw, *Industrial Radiology: Theory and Practice*, 2d ed., Springer, 1995; J. C. Malott and J. Fodor, *The Art and Science of Medical Radiography*, 7th ed., Mosby-Year Book, 1993; K. N. Prasad, *Handbook of Radiology*, 2d ed., CRC Press, 1995; J. Selman, *Fundamentals of X-ray and Radium Physics*, 8th ed., C. C. Thomas, 1994.

X-rays

X-rays, or Roentgen rays, are electromagnetic waves; they are the same as visible light, except that they have shorter wavelengths (higher photon energies). Thus x-rays, visible light, ultraviolet, infrared, microwaves, and radio waves are all electromagnetic radiation in different wavelength (energy) spectral regions. X-rays are generated when fast-moving electrons slow down and stop in matter, when an inner-shell vacancy in an atom is filled by another electron, and when electrons moving at relativistic speeds (speeds near the speed of light) change their direction of motion in space. *See* ELECTROMAGNETIC RADIATION.

Roentgen's findings. X-rays were discovered by W. C. Roentgen in 1895. This discovery came about by accident. Roentgen was studying gas discharges in his laboratory when he noticed that unknown radiation from a gas discharge could induce fluorescence in certain materials. In his first communication, Roentgen described the properties of these rays as follows: They were invisible; moved in straight lines; were unaffected by electric or magnetic fields, and hence not electrically charged; passed through matter opaque to ordinary light (since they penetrated through the black cardboard around his cathode-ray tube); were differentially absorbed by matter of different densities or of different atomic weights; affected photographic plates; produced fluorescence in certain chemicals, such as in the barium platinocyanide screen with which the initial discovery was made and in the wall of his glass tube opposite the cathode; produced ionization in gases; and were evidently produced at the anode by the beam of rays (identified by J. J. Thomson in 1897 as electrons) issuing from the cathode in his vacuum tube. *See* CATHODE RAYS; FLUORESCENCE.

Along with all these definitive characteristics of the rays, however, other crucial experiments designed to establish similarity or differences from visible light were clearly called for. The fundamental optical properties of visible light were well established in 1895: reflection from mirrors; refraction in prisms (change in direction in passing from air into glass, for example), by means of which a beam of white light could be spread out into a rainbow or spectrum of colors; diffraction by narrow slits or ruled gratings, also a method of producing spectra; and polarization, or constraint of the electric field of the light wave to a single direction. In spite of the best efforts of Roentgen, no evidence of any of these four optical phenomena could be found. Hence the designation "x"—unknown—was assigned by Roentgen. Many theories were proposed to account for the apparently unique quality of x-rays, which seemed to be so closely similar and yet so greatly different from visible light. *See* DIFFRACTION; LIGHT; POLARIZATION OF WAVES; REFLECTION OF ELECTROMAGNETIC RADIATION; REFRACTION OF WAVES.

Later discoveries. Other scientists studying x-rays found the essential experimental conditions to prove that x-rays can be polarized (C. Barkla, 1905, by scattering from carbon); diffracted by crystals (M. von Laue, W. Friedrich, and P. Knipping, 1912); refracted in prisms and in crystals; reflected by mirrors; and diffracted by ruled gratings (A. Compton, 1921–1922). Instead of being refracted in passing from a less dense medium (air) to a more dense medium (a glass prism or a crystal) in the same direction as light (the index of refraction for visible light is always greater than 1), x-rays are deviated in the opposite direction by a very small amount: the index of refraction is less than 1 by an amount as small as 10^{-6}. Total reflection from mirrors is observed only when the beam impinges at a very small angle (grazing angle), a necessary condition understandably missed by Roentgen. Similarly, the beam must be incident at a very small angle on a ruled diffraction grating if diffraction (a spectrum, or spatial dispersion by wavelength) is to be observed. *See* X-RAY OPTICS.

From 1895 to 1912 there seemed to be no analyzer capable of dispersing an x-ray beam into a spectrum. The spectacular Laue diffraction pattern of a zinc sulfide crystal in 1912 proved the electromagnetic wave nature of x-rays and the ordered structure of crystals, with atoms lying on families of planes to constitute three-dimensional diffraction gratings, all governed by the simple Bragg law $n\lambda = 2d \sin \theta$ (which must be corrected for refraction in extremely accurate work). Here n is an integer indicating the order of the spectrum, λ the wavelength, d the crystal lattice spacing of one set of planes, and θ the angle between the incident ray and this set of planes. *See* CRYSTAL STRUCTURE; X-RAY DIFFRACTION.

The wavelength range of x-rays in the electromagnetic spectrum, as excited in x-ray tubes by the bombardment of a target by electrons under a high accelerating potential, overlaps the ultraviolet range on the order of 100 nanometers on the long-wavelength side, and the shortest-wavelength limit moves downward as voltages increase. An accelerating potential of 10^9 V, now readily generated, produces a λ of 10^{-6} nm. An average wavelength used in research is 0.1 nm, or about 1/6000 the wavelength of yellow light. *See* ULTRAVIOLET RADIATION; X-RAY TUBE.

Quantum mechanics. X-rays (and visible light) can be considered as an electromagnetic wave. X-rays can also be considered as discontinuous bundles of energy, or quanta, in accordance with the laws first enunciated by M. Planck and extended by A. Einstein early in the twentieth century. In diffraction, refraction, polarization, and interference phenomena, x-rays, together with all other electromagnetic radiation, appear to act as waves and λ has a real significance. There is duality, meaning that light and x-rays have both wave and particle properties, although these are generally not observed at the same time in a given experiment. Beams of electrons and neutrons also have wave properties, and are diffracted in appropriate media. In other phenomena—such as the appearance of sharp spectral lines, a definite short-wavelength limit λ_0 of the continuous "white" spectrum [defined by $\lambda_0 = hc/eV$, where h is Planck's constant, c the velocity of electromagnetic radiation (including light and x-rays), e the charge of electron,

and V the accelerating voltage], the shift in wavelength of x-rays scattered by electrons in atoms (Compton effect), and the photoelectric effect—the energy is propagated and transferred in quanta (called photons) defined by values of hv, where the frequency v is c/λ. *See* COMPTON EFFECT; ELECTRON DIFFRACTION; NEUTRON DIFFRACTION; PHOTOEMISSION; PHOTON; QUANTUM MECHANICS.

Applications. X-rays are a valuable probe of matter, as they interact selectively with electrons; electrons in matter account for most of the significant properties of matter (excepting nuclear properties). Medical uses abound, as x-rays can penetrate matter, and are selectively absorbed by atoms containing many electrons—elements with a high atomic number. Thus dental x-rays show fillings as dark, teeth and bone in grey, and are only lightly absorbed in the cheek. X-rays in high intensities are also used for therapy, to treat certain cancers. X-rays are also used in many industrial processes, as in checking the integrity of welds. Tomographic techniques allow three-dimensional images to be obtained, which are invaluable in medical diagnosis.

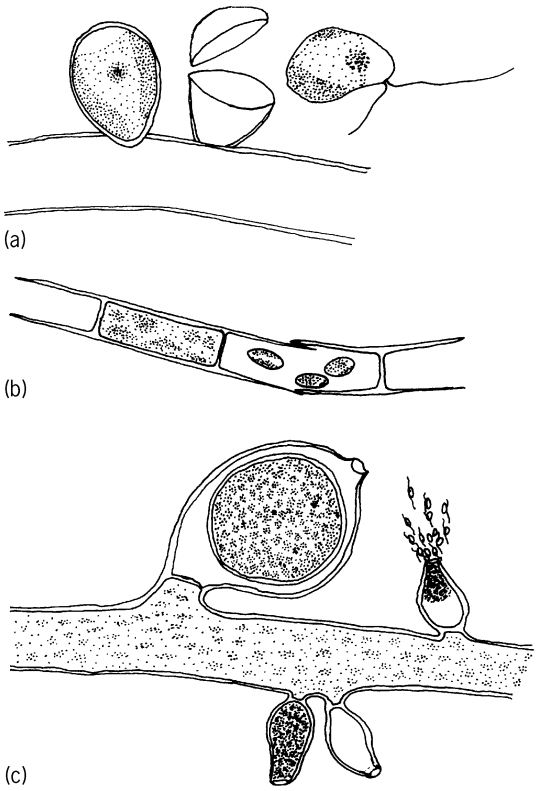
X-rays are used in research for the study of both the electronic structure of matter and its spatial structure. X-ray microscopes and microprobes have been developed which allow imaging with high spatial resolution while obtaining contrast with different elemental discrimination, distinguishing chemical bonds, and, by use of circularly polarized x-rays, the orientation of magnetization. X-rays are used in protein crystallography to determine the spatial structure of proteins, viruses, and other objects that can be made into crystals but that otherwise cannot be seen. *See* COMPUTERIZED TOMOGRAPHY; HISTORADIOGRAPHY; MEDICAL IMAGING; MICRORADIOGRAPHY; NONDESTRUCTIVE EVALUATION; RADIOGRAPHY; RADIOLOGY; X-RAY CRYSTALLOGRAPHY; X-RAY FLUORESCENCE ANALYSIS; X-RAY MICROSCOPE; X-RAY POWDER METHODS. George L. Clark; Alfred S. Schlachter

Bibliography. J. G. Brown, *X-rays and Their Applications*, Plenum, 1975; J. E. Cullinan and A. M. Cullinan, *Illustrated Guide to X-ray Technics*, 2d ed., Lippincott, 1980; N. A. Dyson, *X-rays in Atomic and Nuclear Physics*, 2d ed., Cambridge University Press, 1990; H. Haken and H. C. Wolf, *The Physics of Atoms and Quanta: Introduction to Experiments and Theory*, 7th ed., Springer, 2005; J. Selman, *Fundamentals of X-ray and Radium Physics*, 8th ed., C. C. Thomas, 1994.

Xanthophyceae

A class of plants, comprising the yellow-green algae, in the chlorophyll *a–c* phylectic line (Chromophycota). Alternate names are Tribophyceae, derived from *Tribonema*, a filamentous member of the class, and Heterokontae, referring to the presence of two kinds of flagella on each motile cell. *See* ALGAE; CHROMOPHYCOTA.

Characteristics. The primary photosynthetic pigments are chlorophyll *a* and *c*. Carotenoids include β-carotene and several xanthophylls, among which heteroxanthin and vaucheriaxanthin ester are unique to the class, while diatoxanthin and diadinoxanthin are found also in diatoms and dinoflagellates. As in other Chromophycota, the chloroplasts of yellow-green algae are bounded by a double-membrane envelope surrounded by a chloroplast endoplasmic reticulum, the outer membrane of which is continuous with the outer membrane of the nuclear envelope. Each photosynthetic lamella comprises three thylakoids. Some chloroplasts have a multilamellar internal pyrenoid. Nonpigmented xanthophytes are rare.



Xanthophytes. (*a*) *Lutherella*, a coccoid xanthophyte epiphytic on another alga, showing vegetative cell, an empty two-piece cell wall, and a zoospore. (*b*) *Tribonema*, a filamentous xanthophyte, showing immature zoospores and H-shaped wall segments. (*c*) *Vaucheria*, a siphonous xanthophyte; monoecious species with mature oogonium and three stages of development of antheridium, the mature one releasing antherozoids.

The cell wall is often visibly lamellate, but its chemical composition is poorly known. Cellulose, pectic compounds, and silica are often suggested. Cellulose is known for certain to be the chief component in *Vaucheria*. In many filamentous and some large coccoid forms, the wall of each cell is composed of two overlapping pieces (in filaments, H-shaped in optical section) [**illus.** *a* and *b*]. Endogenous cysts (statospores) also have walls with two equal or unequal pieces. The chemical composition of the food reserves is also poorly known, but leucosin (a β-1,3 glucan) and lipids are usually reported. The absence of starch is an important characteristic in distinguishing yellow-green algae from green algae, which they may resemble superficially in form and color.

Motile cells have two unequal flagella borne apically or laterally (illus. *a*). One (usually the longer) is directed forward and is pleuronematic (hairy); the other is directed backward and is acronematic (smooth). The compound zoospore (synzoospore) of *Vaucheria* is exceptional in having numerous pairs of slightly unequal smooth flagella. There are usually two chloroplasts in each motile cell, the ventral one with a reddish refractile eyespot at its anterior end. One or two contractile vacuoles occur at the anterior end of a motile cell.

Cells in most Xanthophyceae are uninucleate, but in some genera they are multinucleate. The mitochondria have tubular cristae. Each vegetative cell contains one to many yellow-green, laminate or discoid, parietal chloroplasts.

Reproduction. Formation of autospores (nonmotile spores that are miniatures of the cells that produce them), and less often zoospores, is the most common reproductive process. Statospores (internally formed resting stages) are formed by some genera. These germinate to form zoospores or ameboid cells. Sexual reproduction is rarely observed except in *Vaucheria*, which is oogamous. Isogamy and anisogamy are known for a very few genera.

Classification. About 600 species in 115 genera are recognized and distributed among six orders based on the form of the thallus. By far the largest order is Mischococcales, which comprises nonmotile coccoid unicells (illus. *a*). Families and genera within this order are distinguished primarily on the basis of habit (whether the cells are solitary or colonial, attached or free-living, sessile or stipe-borne on a supporting stalk, and with or without mucilage), cell shape, and cell wall ornamentation. Filamentous yellow-green algae constitute the order Tribonematales. *Tribonema* (illus. *b*) is a common inhabitant of standing waters, especially in spring.

Siphonous yellow-greens are placed in the order Vaucheriales. *Vaucheria* (illus. *c*) is the best-known xanthophyte, commonly forming extensive growths on moist soil or mud. Its thallus consists of branched coenocytic filaments with numerous nuclei and chloroplasts concentrated in a peripheral cytoplasmic layer. Septa are formed only to set apart the reproductive structures. Each oogonium contains an egg, while an antheridium produces numerous small colorless antherozoids. Each antherozoid has a pair of laterally inserted flagella, the anteriorly directed one being the shorter, contrary to the usual situation in xanthophyte motile cells. An antherozoid swims through a pore in the oogonial wall to fertilize the egg. The resulting zygote secretes a thick wall and remains in place as a zygospore until the oogonial wall decays. Following a period of dormancy, the zygospore germinates meiotically to produce haploid filaments. Asexual reproduction is effected by multinucleate structures set apart by septa. In terrestrial forms these structures are released as aplanospores, while in aquatic forms they develop a clothing of flagella before being released as compound zoospores.

The orders Chloramoebales, Rhizocloridales, and Heterogloeales comprise flagellate unicells (monads), pseudopodial forms, and palmelloid forms, respectively.

Distribution. Most Xanthophyceae occur in freshwater, especially soft water. Frequently, they are epiphytic on aquatic plants. The chief exception is *Vaucheria*, which has freshwater, brackish-water, and marine species. Xanthophyceae also occur in and on soil and mud, in snow and ice, and on tree trunks and damp walls. Paul C. Silva; Richard L. Moe

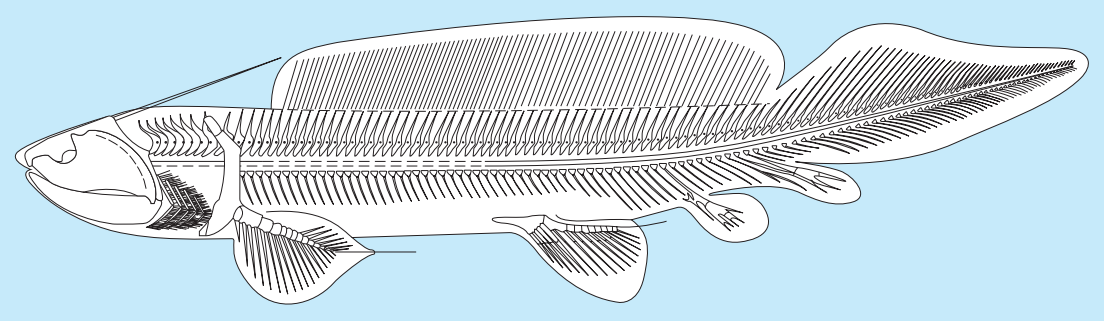
Bibliography. E. R. Cox (ed.), *Phytoflagellates*, pp. 243–271, 1980; H. Ettl, J. Gerloff, and H. Heynig (eds.), *Süsswasserflora von Mitteleuropa*, vols. 3 and 4, 1978.

Xenacanthidae

A family of Paleozoic elasmobranch sharks that are characterized by teeth with large divergent lateral cusps (diplodont crowns), an elongate dorsal fin preceded by a distinctive spine, and paired fins with an unusually well-developed skeletal axis. The xenacanths are one of the most clearly defined and stable groups to emerge from the early radiation of cartilaginous fishes (Chondrichthyes), with body fossils known from the early Carboniferous (*Diplodoselache*) through the late Triassic (*Xenacanthus*) [see **illustration**]. Numerous articulated specimens have been discovered, predominantly from the upper Carboniferous and Permian of Europe and North America, wherein xenacanths are a classic feature of faunas associated with brackish and freshwater lagoons and coal swamps. These xenacanths include the most frequently depicted genera, such as *Orthacanthus* and *Pleuracanthus*, with eellike body forms of up to 3 m (10 ft) or more in length.

Skeletal anatomy. The exceptional abundance and quality of xenacanth fossil material, relative to other Paleozoic chondrichthyans, has allowed their skeletal anatomy to be studied in unusual detail. This has yielded numerous shared, specialized features that ally this group with the lineage leading to modern sharks and rays rather than ratfish (chimaeroids). For a long time these were the only early sharks in which the braincase anatomy was known in any significant detail; likewise for the complex articulations between the jaws and braincase, which are far less mobile than those of modern elasmobranchs. The skeletons of the paired fins in more derived (evolutionarily advanced) members of the Xenacanthidae are noted for their convergent resemblance to the paired fins of lungfish (Dipnoi). In both examples, an elongate, segmented axial rod extends to the fin extremity, with pre- and postaxial branches of the fin skeleton creating a treelike pattern.

Diplodont teeth. The fossil record of xenacanth teeth exceeds that of the less resilient cartilaginous skeleton. Diplodont teeth are among the very earliest examples of sharklike teeth known, and include *Leonodus* from the Early Devonian of Europe. However, the attribution of these teeth to xenacanths is



***Xenacanthus meisenheimensis*, from the Permo-Carboniferous Saar-Nahe Basin, southwestern Germany. (*After U. H. J. Heidtke, 2003*)**

uncertain; it is likely that simple, twin-cusped teeth have a wide distribution among primitive sharklike fishes, extending well beyond the boundaries of the Xenacanthidae. *See* CHONDRICHTHYES; ELASMOBRANCHII.

Michael Coates

Bibliography. J. A. Long, *The Rise of Fishes*, Johns Hopkins University Press, Baltimore, 1995; J. G. Maisey, *Discovering Fossil Fishes*, Westview Press, New York, 1996.

Xenolith

A rock fragment enclosed in another rock, and of varying degrees of foreigness. Cognate xenoliths, for example, are pieces of rock that are genetically related to the host rock that contains them, such as pieces of a border zone in the interior of the same body. Included blocks of unrelated rocks are more deserving of the xenolith label. Such foreign rocks help establish the once fluid and hence molten condition of invading magma capable of incorporating and mixing an assemblage of unrelated rock inclusions, as at Hutton's Rock near Edinburgh, Scotland. Dark inclusions, commony called enclaves, up to many feet in length can be found in granitic plutons and silicic lava domes and flows. In many cases, these have fine-grained borders against the enclosing rock and are interpreted to be lenses of formerly fluid basaltic magma which crystallized as a result of contact with cooler surrounding silicic magma.

Xenoliths tend to react with the enclosing magma, so that their constituent minerals become like those in equilibrium with the melt. Reaction is rarely complete, however. Even completely equilibrated xenoliths may be conspicuous because the equilibration process does not require either the texture or the proportions of the minerals in the xenolith to match those in the enclosing rock. *See* PLUTON.

Xenoliths may be angular to round, millimeters to meters in diameter, aligned or haphazard, and sharply or gradationally bounded (**Figs. 1** and **2**). Xenoliths are present in most bodies of igneous rock, although they are rare in many igneous rocks. Many rocks are inferred to be igneous because they contain xenoliths, the incorporation of which is taken as evidence of fluidity of the host. Ambiguity arises in the case of some sedimentary rocks such as conglomerate and tillite which may become so thoroughly recrystallized that the sedimentary rock fragments appear to be xenoliths surrounded by crystalline magma.

Some remarkable xenoliths include the metallic iron-rich rock fragments found in some basaltic rocks which have intruded carbon-rich sedimentary rocks and the peridotite xenoliths found in diamond pipes and in some basaltic rocks. Scientists infer some of the peridotite xenoliths to be actually pieces of the

Fig. 1. Xenoliths in granodiorite, Sierra Nevada, California, appear soaked and drawn out parallel to the hammer handle. Hammer is 10 in. (25 cm) long. (*W. B. Hamilton, USGS*)

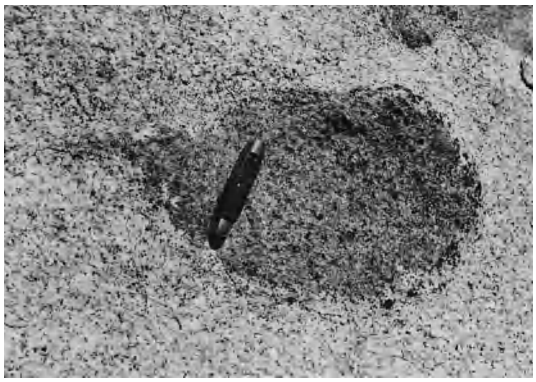


Fig. 2. Xenolith in granite, Sierra Nevada, California. Note sharper contact on one side and gradational contact against contaminated granite on other. (*W. B. Hamilton, USGS*)

Earth's mantle. *See* IGNEOUS ROCKS; LAVA; MAGMA; PERIDOTITE. Alfred T. Anderson, Jr.

Bibliography. J. E. Nielson and J. K. Nakata, Mantle Origin and Flow Sorting of Megacryst-Xenolith Inclusions in Mafic Dikes of Black Canyon, Arizona, *USGS Prof. Pap.*, no. 15, 1994; S. Shi et al., Xenolith evidence for lithospheric melting above anomalously hot mantle beneath the northern Canadian Cordillera, *Contrib. Mineral. Petrol.*, 131:39–53, 1998.

Xenon

A chemical element, Xe, atomic number 54. It is a member of the family of noble gases, group 18 in the periodic table. Xenon is colorless, odorless, and tasteless; it is a gas under ordinary conditions (see **table**). *See* INERT GASES; PERIODIC TABLE.

Xenon is the only one of the nonradioactive noble gases which forms chemical compounds that are stable at room temperature. Xenon also forms weakly bonded clathrates with such substances as water, hydroquinone, and phenol. *See* CLATHRATE COMPOUNDS.

The three fluorides, XeF_2, XeF_4, and XeF_6, are thermodynamically stable compounds at room temperature, and they may be prepared simply by heating mixtures of xenon and fluorine at 300–400°C (570–750°F).

The reaction of XeF_6 with water gives $XeOF_4$; if the reaction is allowed to continue, XeO_3 is formed. XeO_3 is a colorless, odorless, and dangerously explosive white solid of low volatility. Gaseous xenon tetroxide, XeO_4, is formed by the reaction of sodium perxenate, Na_4XeO_6, with concentrated H_2SO_4. The vapor pressure of XeO_4 is about 3.3 kPa at 0°C (32°F). It is unstable and has a tendency to explode.

Xenon is produced commercially in an air-separation plant. The air is liquefied and distilled. The oxygen is redistilled; the least volatile portion contains small amounts of xenon and krypton, which are adsorbed on silica gel directly from the liquid oxygen. The crude xenon and krypton thus obtained are separated and further purified by distillation and selective absorption, at controlled low temperatures, on activated carbon. Remaining impurities are removed by passing the xenon over hot titanium, which reacts with all but the inert gases. *See* AIR SEPARATION.

Physical properties of xenon

Property	Value
Atomic number	54
Atomic weight (atmospheric xenon only)	131.30
Melting point (triple point)	−111.8°C (−169.2°F)
Boiling point at 1 atm pressure	−108.1°C (−162.6°F)
Gas density at 0°C and 1 atm pressure, g/liter	5.8971
Liquid density at its boiling point, g/ml	3.057
Solubility in water at 20°C, ml xenon (STP) per 1000 g water at 1 atm partial pressure of xenon	108.1

Xenon is used to fill a type of flashbulb used in photography and called an electronic speed light. These bulbs produce a white light with a good balance of all the colors in the visible spectrum, and can be used 10,000 times or more before burning out.

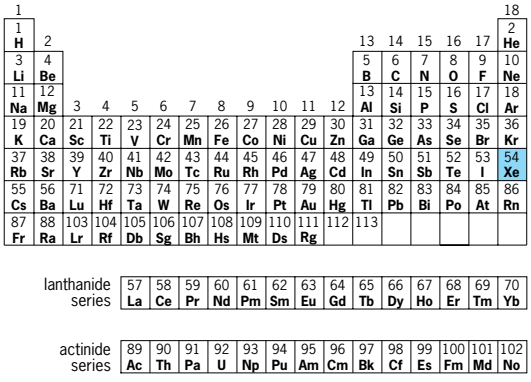
A xenon-filled arc lamp gives a light intensity approaching that of the carbon arc; it is particularly valuable in projecting motion pictures. *See* VAPOR LAMP.

An important development in high-energy physics was the detection of nuclear radiation, such as gamma rays and mesons, by bubble chambers, in which a liquid is kept at a temperature just above its boiling point. Nucleation by the radiation results in bubble formation along the path of the particle. The tracks made by the particles are then photographed. Liquid xenon is one of the liquids used in these bubble chambers.

Xenon is used to fill neutron counters, x-ray counters, gas-filled thyratrons, and ionization chambers for cosmic rays; it is also used in high-pressure arc lamps to produce ultraviolet radiation.

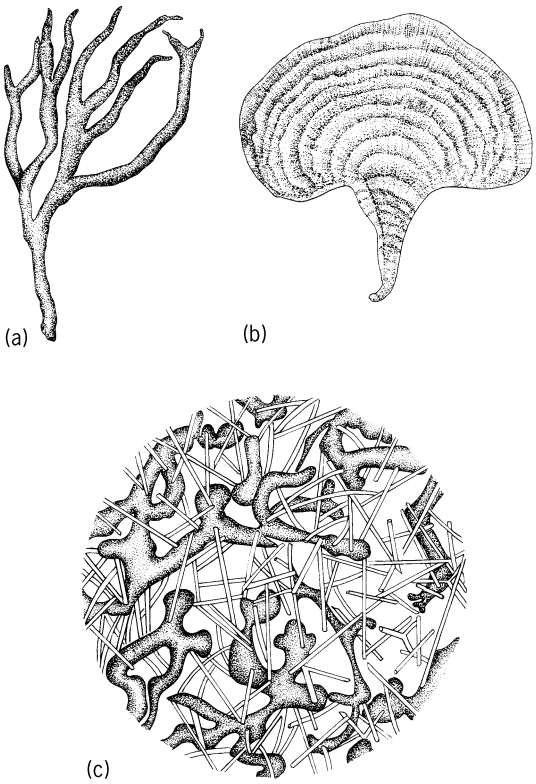
Between 3 and 5% of the fissions in a nuclear reactor using uranium as fuel lead to the formation of xenon-135. Arthur W. Francis

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., CRC Press, 2004; M. Ozima and F. A. Podosek, *Noble Gas Geochemistry*, 2001.



Xenophyophorida

An order of Protozoa in the subclass Granuloreticulosia. The group includes deep-sea forms which are multinucleate at maturity. They develop as discoid to fan-shaped or algalike branching forms (**illus.** *a* and *b*) covered with a hyaline organic layer and sometimes measuring 0.08 in. (2 mm) or more overall. The aggregate contains many tubes (illus. *c*) which vary in diameter and form netlike or branching patterns. The more delicate tubes contain multinucleate protoplasm; the coarser ones, dark brownish clumps which are possibly waste



Xenophyophorids. (*a*) *Stannoma dendroides*. (*b*) *Stannophyllum zonarium*. (*c*) *Psammetta erythrocytomorpha*, section showing interlacing tubules and dark inclusions of uncertain nature.

materials. Foreign particles, such as radiolarian skeletons or foraminiferid tests, are often embedded in the nets or among the branches. A feltlike material also is found between adjacent tubes in some species. The lack of detailed information makes the taxonomic status of these organisms somewhat uncertain. The genera included are *Psammetta*, *Stannoma*, and *Stannophyllum*. *See* GRANULORETICULOSIA; PROTOZOA; RHIZOPODEA; SARCODINA; SARCOMASTIGOPHORA.

Richard P. Hall

Xenoturbella

A simple marine worm whose natural habitat is in soft mud bottoms at a depth of around 60 m (200 ft) off the coasts of Sweden and Norway. *Xenoturbella* was first discovered in 1915 by the Swedish biologist Sixten Bock, but was not described until 30 years later by Einar Westblad. Due to its simplicity of form, Westblad initially classified *Xenoturbella* as a primitive turbellarian, a class of ciliated nonparasitic flatworms, and named it accordingly (*Xenoturbella* = strange flatworm). However, recent analyses using deoxyribonucleic acid (DNA) have found *Xenoturbella* to be a primitive member of the deuterostomes, the group of animals that includes the vertebrates.

Morphology. *Xenoturbella* is a small worm up to 3 cm (1.2 in.) in length with a mouth opening into a gut cavity (**Fig. 1**). It has no anus and lacks defined excretory structures, body cavities, and reproductive organs. It has a gravity-sensing organ called a statocyst and two sensory grooves with a dense underlying nerve plexus. *Xenoturbella* has neither brain nor condensed nerve cord; its nervous system consists simply of a nerve net (a diffuse network of neurons) under the epidermis. Externally, it is completely covered in hairlike structures called cilia, which it uses to glide along the muddy seafloor. The reproduction and embryology of Xenoturbella are still a mystery.

Phylogeny. *Xenoturbella* has been the subject of controversy for almost a century. Despite numerous attempts at assigning *Xenoturbella* a position in the evolutionary tree of life, this animal has remained enigmatic. The problem with finding out its phylogenetic affinities (its relationship to other groups of organisms) has been its extreme morphological simplicity.

Following its initial classification, *Xenoturbella* was believed to be the most primitive bilaterally symmetrical animal. Another hypothesis supported a relationship with the hemichordates (acorn worms) and echinoderms (starfishes and sea urchins), based on ultrastructural similarities of the epidermis, the presence of a nerve net, and similarities between the statocysts of *Xenoturbella* and those of a group of echinoderms, the sea cucumbers.

In 1997, a molecular study based on the analysis of ribosomal DNA and eggs found inside the animal seemed to suggest a relationship to bivalve mollusks of the genus *Nucula*. The fact that nuculid bivalves are abundant in the same habitat as *Xenoturbella* suggested instead that these results might be contaminated with food ingested by *Xenoturbella* in the form of mollusk eggs and embryos.

More recent molecular analyses have revealed that *Xenoturbella* belongs within the deuterostomes. The deuterostomes contain the chordates (the phylum to which vertebrates and humans belong), the hemichordates (acorn worms), and the echinoderms (starfishes). Further analyses of the genetic code and the order of genes on the mitochondrial genome of *Xenoturbella* suggest a placement at the base of hemichordates and echinoderms (**Fig. 2**). These analyses suggest that *Xenoturbella* is a member of a separate phylum, the *Xenoturbellida*.

Because of the pivotal phylogenetic position of *Xenoturbella* as the closest sister group to the hemichordates and echinoderms and an outgroup to

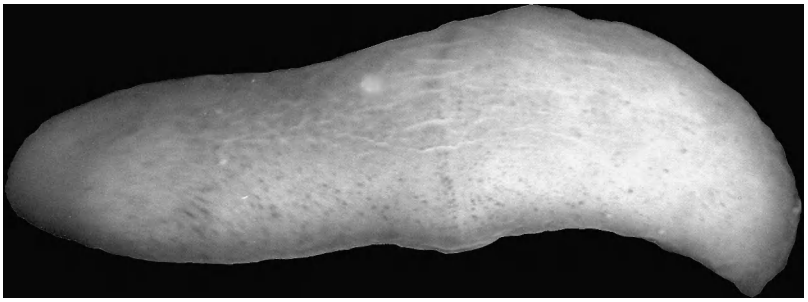


Fig. 1. Photograph of a live specimen of *Xenoturbella* (1–2 cm), showing the sensory circumferential groove around its middle.

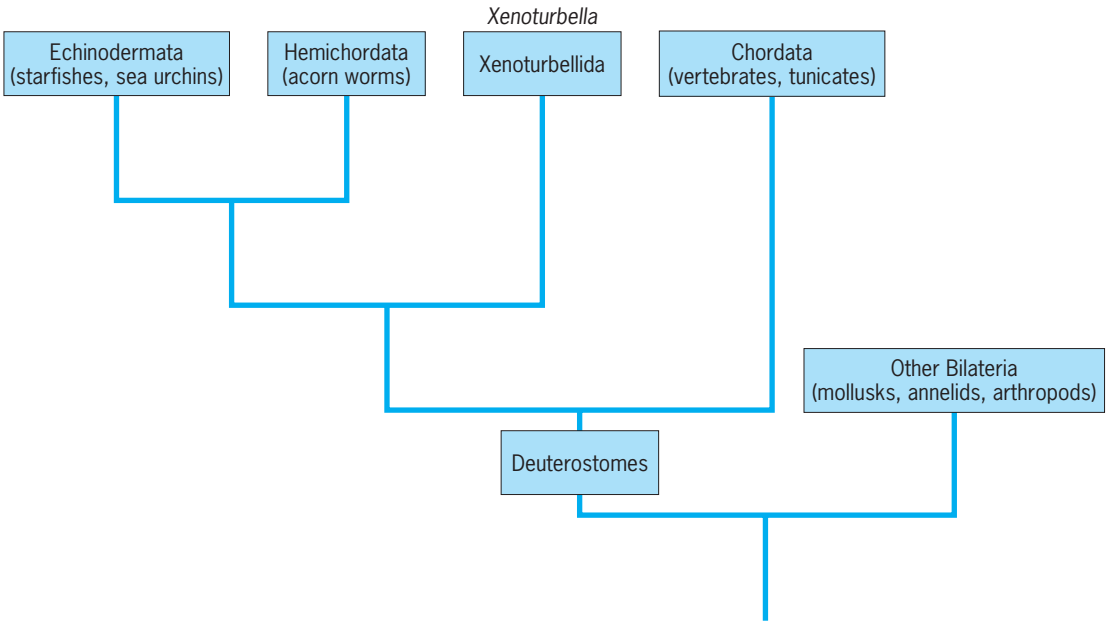


Fig. 2. Phylogenetic tree based on the 18s ribosomal DNA gene, showing the position of *Xenoturbella* at the base of the hemichordates and echinoderms.

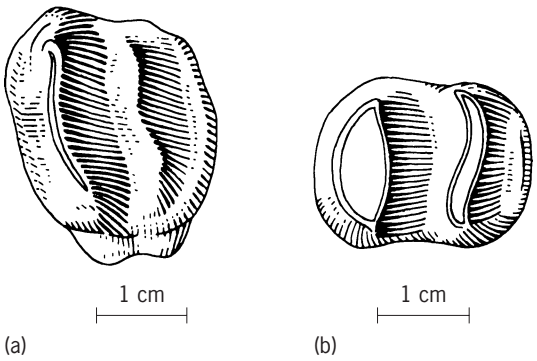
the chordates, future studies of the genetics, morphology, and embryology of *Xenoturbella* have the potential to provide great insight into the evolution of the deuterostomes. *See* ANIMAL EVOLUTION; ANIMAL SYSTEMATICS; DEUTEROSTOMIA; ECHINODERMATA; GENE AMPLIFICATION; HEMICHORDATA; PHYLOGENY.

Sarah J. Bourlat

Bibliography. S. J. Bourlat et al., *Xenoturbella* is a deuterostome that eats molluscs, *Nature*, 424:925–928, 2003; H. Gee, You aren't what you eat, *Nature*, 424:885–886, 2003.

Xenungulata

Xenungulates are large, primitive, digitigrade, hoofed, tapirlike mammals with relatively short, slender limbs and five-toed feet with broad, flat phalanges. They are restricted to the Paleocene deposits of Brazil and Argentina. Only one genus (*Carodnia*) constitutes the single family (Carodniidae) in the order. *See* EUTHERIA.



Molars of *Carodnia vieiral* from the Paleocene of Brazil: (*a*) upper and (*b*) lower.

The dentition (I 3/3 C 1/1 Pm 4/4 M 3/3) is complete with strong, procumbent, chisel-shaped incisors, strong sharp-pointed canines, and low-crowned cheek teeth with bilophodont molars (see **illus.**). *See* DENTITION.

The affinities of the Xenungulata remain uncertain. The combination of the dental characters is unique among the South American ungulates. Affinities with the Dinocerata are strongly supported by the dental characteristics, but the structure of the tarsus suggests that the xenungulates had common ancestry with the Pyrotheria, to which they were tentatively assigned originally. *See* DINOCERATA; PYROTHERIA.

Gideon T. James; Everett C. Olson

Bibliography. M. J. Benton, *Vertebrate Paleontology*, 1991; R. Carroll, *Vertebrate Paleontology and Evolution*, 1988.

Xiphosurida

An order (xiphosurids, or horseshoe crabs) of the class Xiphosura (phylum Arthropoda). These "sword-tailed" arthropods are a classic example of a "living fossils" group that has seemingly changed little over a period of many millions of years.

Xiphosurids have three distinct body divisions: a prominent horseshoe-shaped, fused head and thorax, called the prosoma; a fused "abdomen," or opisthosoma; and a spikelike terminal segment, called the telson (**Fig. 1**). Xiphosurida is a low-diversity group that has a moderately good fossil record in some paleoenvironments, especially marginal-marine environments. Present-day xiphosurid species are tolerant of wide ranges of salinity, from marine to brackish water, and occasionally even venture into freshwater; ancient species may have

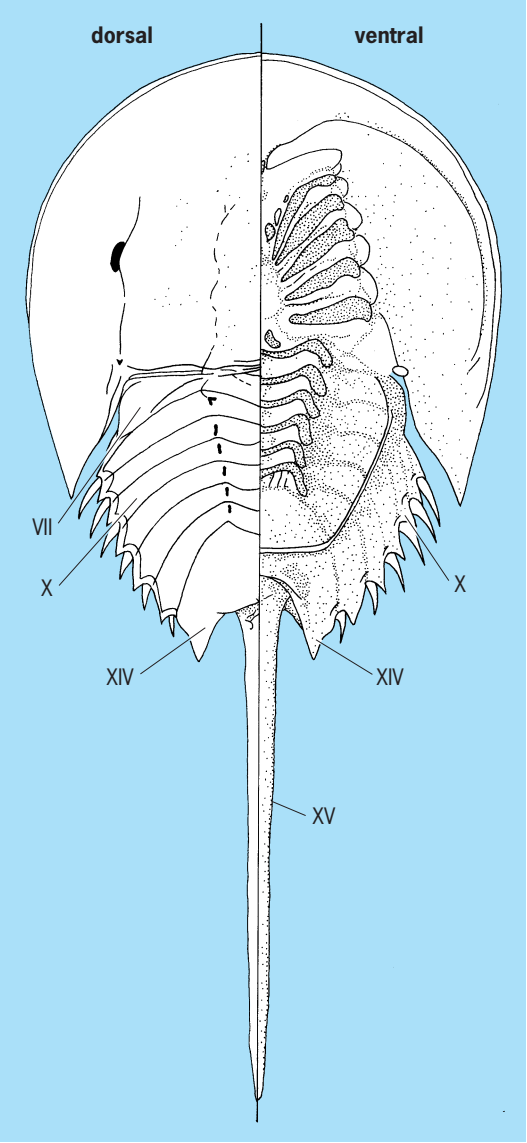


Fig. 1. Segmentation of *Limulus polyphemus*. Inconclusive evidence indicates that there are 15 segments (6 in the prosoma, 9 in the opisthosoma). The telson (XV) may be a terminal spine that has become articulated and has incorporated the anus into its base.

had similar salinity tolerances.

The order Xiphosurida is a monophyletic group of arachnomorph arthropods, a large group that includes nektaspids, trilobites, eurypterids, arachnids, and synziphosurines. Two xiphosurid suborders are recognized: the more primitive Bellinurina, all of which are now extinct, and the more derived Limulina, which includes all the four living species of the group (**Fig. 2**). The most primitive group of xiphosurids (suborder Bellinurina) probably evolved from a stem-group xiphosuran, called a synziphosurine, during the Late Devonian (about 365 million years ago), and became extinct in the Permian Period. In the synxiphosurines, which are the sister group to the bellinurine xiphosurids, the segments of the opisthosoma articulate freely and the telson is short and usually wide. Some species lack compound eyes. Evolution of the bellinurines involved fusion of the opisthosoma into a single plate sometimes called a thoracetron, and an increase in the length of the telson. Other developments include the addition of true lateral spines on the opisthosoma, and development of a small node bearing light-sensitive eyes (ocelli) in front of the axial, or cardiac, lobe on the prosoma. The Limulina evolved from a bellinurine ancestor in the early Carboniferous (about 335 million years ago). The most important evolutionary innovation in the limulines is the change in shape of the small lobe at the rear of the opisthothorax from triangular (the condition in bellinurines) to trapezoidal. Limulines are also characterized by having lost the axial part of the opercular tergite (the eighth tergite) and by the formation of conical projections bearing the genital pores on the back of the operculum. *See* MEROSTOMATA.

There are four extant species of limulines: *Limulus polyphemus*, which occurs on the eastern coast of North America, and three species distributed in the Indo-Pacific region (*Tachypleus tridentatus, T. gigas*, and *Carcinoscorpius rotundicauda*). Phylogenetic studies based on mitochondrial genes suggest that *L. polyphemus* is the sister group to all of the Indo-Pacific species, and that *T. gigas* and *C. rotundicauda* form a monophyletic group that has the less derived *T. tridentatus* as its sister group.

Fossil record. Xiphosurids include 13 genera known only from fossils, two genera known only from living species, and one genus that is known from both fossil and living species. The suborder Bellinurina includes the Late Paleozoic genera *Bellinurus, Liomesaspis, Euproops, Bellinuroopsis*, and *Rolfia*. The suborder Limulina includes the Late Paleozoic genera *Xaniopyramus, Paleolimulus, Panduralimulus*, and *Valloisella*; the Mesozoic genera *Limulitella, Psammolimulus*, and *Mesolimulus*; the Late Mesozoic-Cenozoic genus *Limulus*; and the exclusively Cenozoic genera *Tachypleus* and *Carcinoscorpius*.

Xiphosurids seem to have an anomalous fossil record. Their exoskeletons, like those of most arthropods, are nonbiomineralized; they are composed of chitin, which is a polysaccharide that does not normally fossilize. Moreover, fossil xiphosurids are unusual because they are commonly found complete, retaining the prosoma, the opisthothorax, and the telson. In some instances, ventral appendages and book gills are also attached (**Fig. 3**).

Studies on the fossilization history (taphonomy) of modern *Limulus* have elucidated some of the reasons that xiphosurids have such an unusual fossil record, making them a model for the preservation of other types of nonmineralizing arthropods. *Limulus* disarticulates slowly after death. As long as the bodies remain in water, smaller specimens in particular tend to become pliable, and subject to wrinkling and compression when buried under sediment. Normally, it takes at least 2 weeks for the telson to separate from the opisthosoma, and 4 weeks or more for the opistosoma to separate from the prosoma. The slow rate of separation of the major body divisions helps account

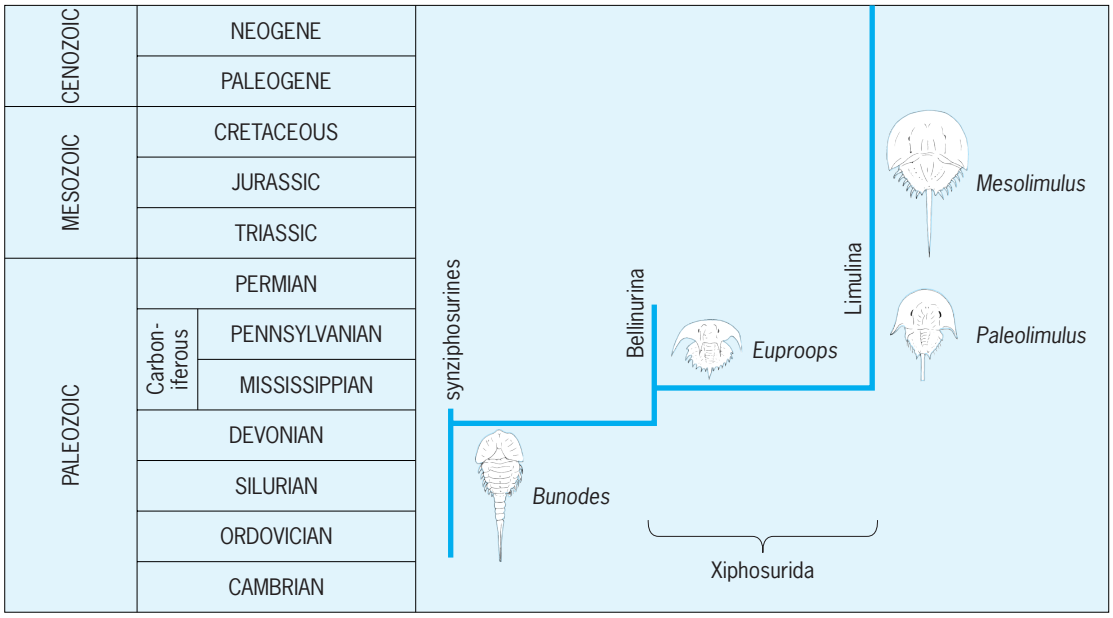


Fig. 2. Phylogeny and representive species of the class Xiphosura, including the order Xiphosurida and its sister group, synziphosurines (a paraphyletic stem group).

for the common occurrence of complete horseshoe crab fossils.

Within a few days after death, modern horseshoe crabs become enveloped in a thin film or "halo" composed of microscopic fungi and bacteria. The microbes apparently aid in decay of the chitinous exoskeleton, but at some point bacteria seem to initiate the process of fossilization of the exoskeleton by causing precipitation of minerals such as calcite (calcium carbonate) or siderite (iron carbonate) on the surface of the horseshoe crab. Precipitation of minerals on the exoskeleton apparently can begin within a few weeks of death, so a slow disarticulation rate coupled with a rapid fossilization rate contributes to the unusually good fossil record of these animals.

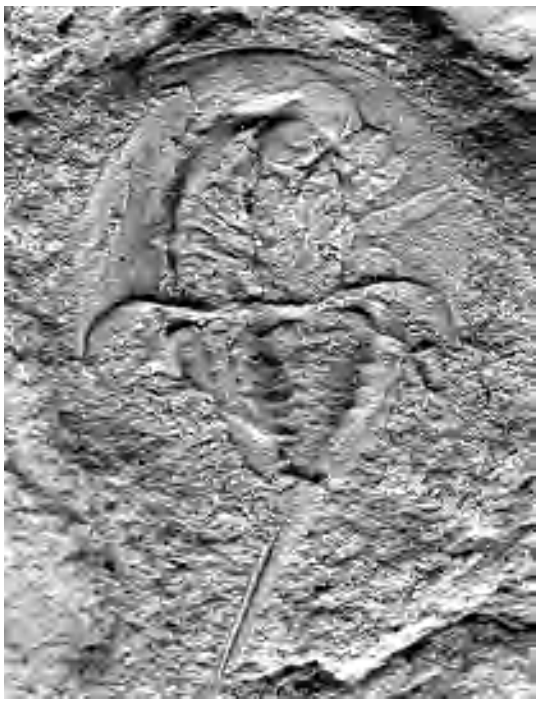


Fig. 3. *Paleolimulus signatus*, a limuline xiphosurid from the Permian of Kansas, showing appendages and book gills. Length of specimen: 1.25 cm (0.5 in.).

It is also possible that factors related to habitat played an important role in the fossilization of xiphosurids. Often their remains become buried in estuarine, tidal flat, or lacustrine environments. Salinity fluctuations, and possibly desiccation, in these settings may have limited the number of scavengers and sediment burrowers that could destroy xiphosurid remains. *See* TAPHONOMY.

Distribution. Distribution of fossilized and living species suggests that limuline xiphosurids have always been tropico-temperate species, occurring mostly in the Northern Hemisphere. None of the extant species has a continuous distribution; each has an intermittent distribution in coastal waters. Occasionally, individuals are found well upstream from coastal estuaries.

Limulus polyphemus is found along the Atlantic coast of North America from Maine to the Yucatán Peninsula, Mexico. The largest adults and largest populations occur in the middle part of the species range, from Delaware Bay to Georgia.

The ranges of the three species endemic to the Indo-Pacific region overlap in part. *Tachypleus tridentatus* occurs along the western and southern shores of Japan, along the southeastern coast of China to southern Vietnam, and along the western islands of the Philippines. *Tachypleus gigas* occurs along the shores of the Bay of Bengal, and from Indochina to northern Vietnam, Borneo, and the Celebes Islands of Indonesia. *Carcinoscorpius rotundicauda* extends from the western shore of the Bay of Bengal to the southern coast of the Philippines. In addition to geographic ranges of

comparable extent, the four species have similar ecologies, morphologies (**Fig. 4**), and serologies.

Life history. *Limulus polyphemus* is the most studied species of extant xiphosurid. Its common names include horseshoe crab (preferred), horsefoot, pan crab, piggyback crab, and king crab (not to be confused with *Paralithodes camtschatica*, the true king crab).

The characteristics, including behavior, of any one population are not exactly like those of another. There is evidence of genetic variability and, for whatever reason, obvious adjustments to local environments, such as the zone of the beach where eggs are laid: near the high-tide mark where the tidal level varies about 1 m (3.3 ft), as along the Gulf of Mexico coast of Florida; throughout the upper portion of the beach when the range between low and high tide is about 2 m (6.6 ft), as in Delaware Bay; and in the middle portion of the beach when the tidal amplitude exceeds 3 m (10 ft), as in the Cape Cod Bay area.

The life history of *L. polyphemus* in Delaware Bay, where the largest population abounds, is generally characteristic of the species. Adults spend the colder months in deeper waters within the bay and on the continental shelf. The geographic extent of this population is not known, but it may be far-ranging; other *Limulus* to the south, off North Carolina, have been dredged from depths of 66 m (218 ft) some 56 km (35 mi) off the coast. Warming waters and increasing daylight stimulate the adults to migrate toward the sandy bay beaches to spawn. Most follow the path of least resistance and follow the strong flood tide currents into the eastern side of Delaware Bay. Fewer horseshoe crabs gather along the western shore, except when rough waves strike Cape May beaches during prevailing northwesterlies and drive the animals back into the deeper water. Then they may venture westward toward the wind-protected shore.

On new moon, and especially full moon, and during high tides in May and June, thousands of *L. polyphemus* adults congregate on the beaches to spawn. Males "patroling" along the shore grasp the females heading for the beach. In an excavation dug beneath her body, each female lays several thousand eggs that are fertilized when sperm are washed into the nests by the waves. After making one or more nests, the animals leave the beach when the tide ebbs. Horseshoe crabs that do not leave before the tide recedes are stranded on the tidal flats, where they dig in to await the next high water.

Minnows swarm along the beaches at night and feed upon the eggs washed out of the nests by wave action. During the day sharp-eyed shore birds, mostly migrants such as plovers and sandpipers, dig into the nests and gorge themselves on eggs.

Horseshoe crab embryos begin to develop within opaque, tough-coated eggs. In a few days each egg cover ruptures and a thin, transparent sphere, formed during embryonic development, balloons out to assume a diameter larger than the original egg (**Fig. 5***a*). In this transparent showcase, growth of the legs, their first feeble attempts to move, and then their almost incessant movement are clearly visible (Fig. 5*b*).

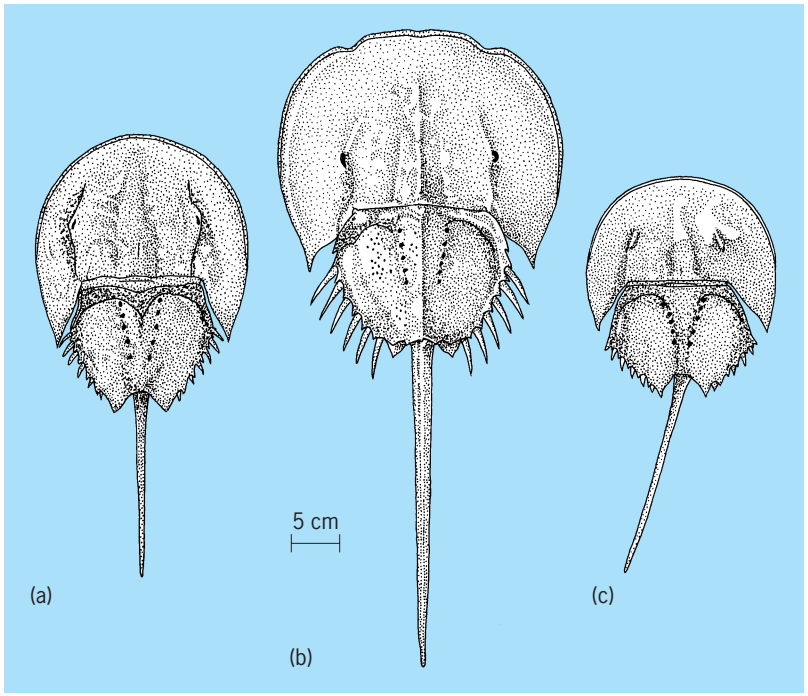


Fig. 4. Carapaces of adult male Limulina. (*a*) *Limulus polyphemus*. (*b*) *Tachypleus tridentatus*. (*c*) *Carcinoscorpius rotundicauda*.

Within 2 weeks some of the tiny animals reach the larval stage and are ready to hatch, although development and hatching may be delayed for several weeks. If the waves churn up a nest, the moving sand grains rupture the enclosing membrane and liberate the larvae (Fig. 5*c*). From then on, until they reach maturity in some 9 to 11 years and 15 or more stages later, their life is essentially seeking food and growing. Food consists mainly of marine worms, such as *Nereis*, and soft-shelled mollusks, such as *Mya arenaria*. The larger the immature animal, the more it moves toward the deeper bay water, from which, as an adult, it begins the spawning migration to renew the life cycle.

The newly hatched larva, which is referred to as the trilobite or Euproops stage, is in many respects still an embryo. Its incompletely formed digestive tract and nervous and circulatory systems develop further during early subsequent stages. The yolk-filled midgut serves as food for the larvae while they learn to right themselves from the upside-down position in which gravity held them during development in the egg (Fig. 5*c*). Horseshoe crabs of all ages swim on their backs, however, and usually near the water surface. *See* INVERTEBRATE EMBRYOLOGY.

Molting. Molting occurs several times during the first 2 or 3 years of life, but rarely more than once thereafter until adult size is reached. It is preceded by the formation of the new skin, recognizable externally in the larger animals by a deeper olive-green coloration with yellow margins of the shell. The new skin forms pleats that unfold as the animal expands with the uptake of water during emergence. These

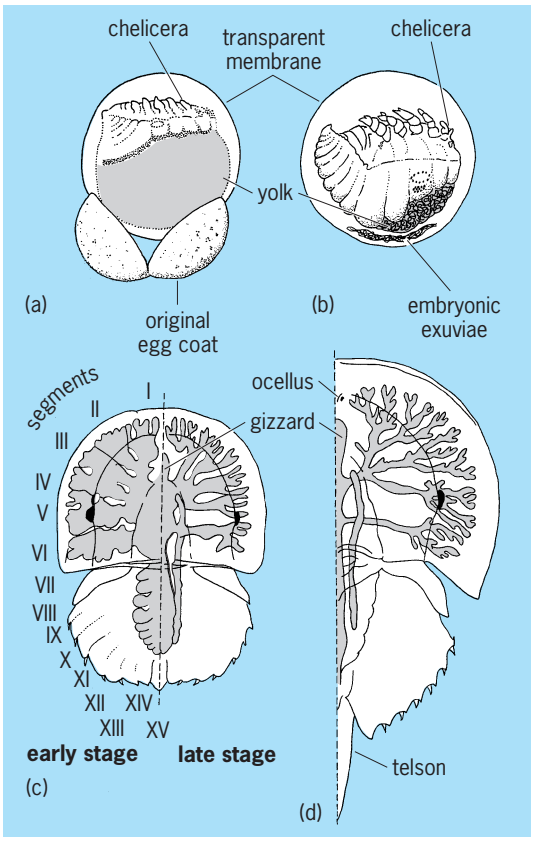


Fig. 5. Embryonic and larval development of *Limulus polyphemus.* (*a*) Early embryo, side view. (*b*) Late embryo. (*c*) Yolk-filled digestive tract of larva. Note the segmentation of yolk in the newly hatched stage and the six lobes of digestive gland. (*d*) First tailed stage.

pleats in the premolt skin give the hardened exoskeleton its dark-lined, mosaic appearance. A circumferential splitting, just within the ventral margin of the prosoma, is the first outward evidence of molting. Exuviation (casting off the shell) then proceeds slowly until the animal emerges one-third of the way and is loosened from the old shell. Emergence then proceeds more rapidly. The first tailed stage emerges within an hour, but an animal 10 cm (4 in.) wide may require more than 24 h to shed its old shell. Young specimens molt earlier than older ones, with cast shells appearing in sequence during July and August on tidal flats and in beach wrack. Instinctively, the larger, annually molting animals seek the areas along the edges of tidal flats and dig in prior to molting. Burying protects them from their enemies: birds, fishes, and crabs. The moist substrate supplies water at all times.

Nutrition. In feeding, the pincer-tipped legs grasp and bring prey to the heavy bases of the pushers, the fifth pair of legs, where shells of mollusks are crushed. A pair of paddlelike appendages and spines on the leg bases then push and knead the prey forward toward the mouth, which is surrounded by the leg bases. Chelipeds in front of the upper lip push food backward. Chitinous ridges of the muscular gizzard, with the aid of ingested sand and shell, grind up prey into pulp that is squeezed into the midgut, where it is partially digested. The resulting chyle is forced into the many-branched digestive gland, where digestion within the cells completes the process. Large pieces of shell and pebbles are regurgitated from the gizzard, but fine sand and smaller shell fragments may be passed with undigested residues in soft, slime-enveloped, fecal cylinders.

Anatomy. The carapace (test or shell) is composed of chitin, strengthened internally in the adults by bridgelike struts within the shell margins (**Fig. 6**). Sensory structures on the carapace include chemoreceptors for taste on the bases of the legs and a pair of simple eyes, or ocelli. The multilensed compound eyes form mosaic images and are sensitive to polarized light.

Adult males are easily distinguished from the larger females by the higher anterior arch of the prosoma, by the pair of "fish-and-thumb" claspers used in mating, and by the hard, conical projections bearing the genital pores on the back of the operculum (eighth appendage). The female apertures are larger, soft, and oval.

The organ systems, except the peripheral blood vessels and nerves, digestive, and reproductive glands, which branch extensively into the lateral portions of the prosoma, are mainly within the prominent axial region (Fig. 6*c*). This region contains the musculature of the paired appendages, hinge, and telson, the digestive tube, the coxal (excretory) gland, and the respiratory, central circulatory, and nervous structures. The dorsally situated tubular heart lies within a sinus that receives venous blood, the hemolymph, from the book gills (the respiratory structures) and the body. The central nervous system and the major nerve branches are sheathed by arteries. A plate of "cartilaginous" chitin, the endocranium, forms a roof over the brain. Flat, elliptical, primitive white blood cells, the amebocytes, circulate in the hemolymph. When exposed to air, the amebocytes send out pseudopods and fragment to form a clot. The blood pigment is hemocyanin. *See* RESPIRATORY PIGMENTS (INVERTEBRATE).

Importance. Early colonists from Europe saw Native Americans use the tails of horseshoe crabs for spearheads and the animals for fertilizer. In the Delaware Bay area, prior to 1960, horseshoe crabs were ground up to form a meal having a protein content of 46% and used in fertilizer. Quartered animals make good eel bait and have been used as food for hogs and poultry. The roe reportedly increases growth of broilers and egg laying, although a fishy flavor is imparted to both flesh and eggs. Eel fishery has made heavy inroads into some *L. polyphemus* populations.

Limulus polyphemus is particularly useful in biomedical research. The nerve and muscle tissue of the heart are easily separable anatomically. All the nerve cells lie in compact masses on the surface of the heart and can be removed without damaging the muscle tissue. Studies on the heart of horseshoe crabs have played an important role in the understanding of cardiac function in vertebrates. *Limulus polyphemus* has been used extensively in research on the visual system. The compound eye polarizes light, and its crystalline cones concentrate light

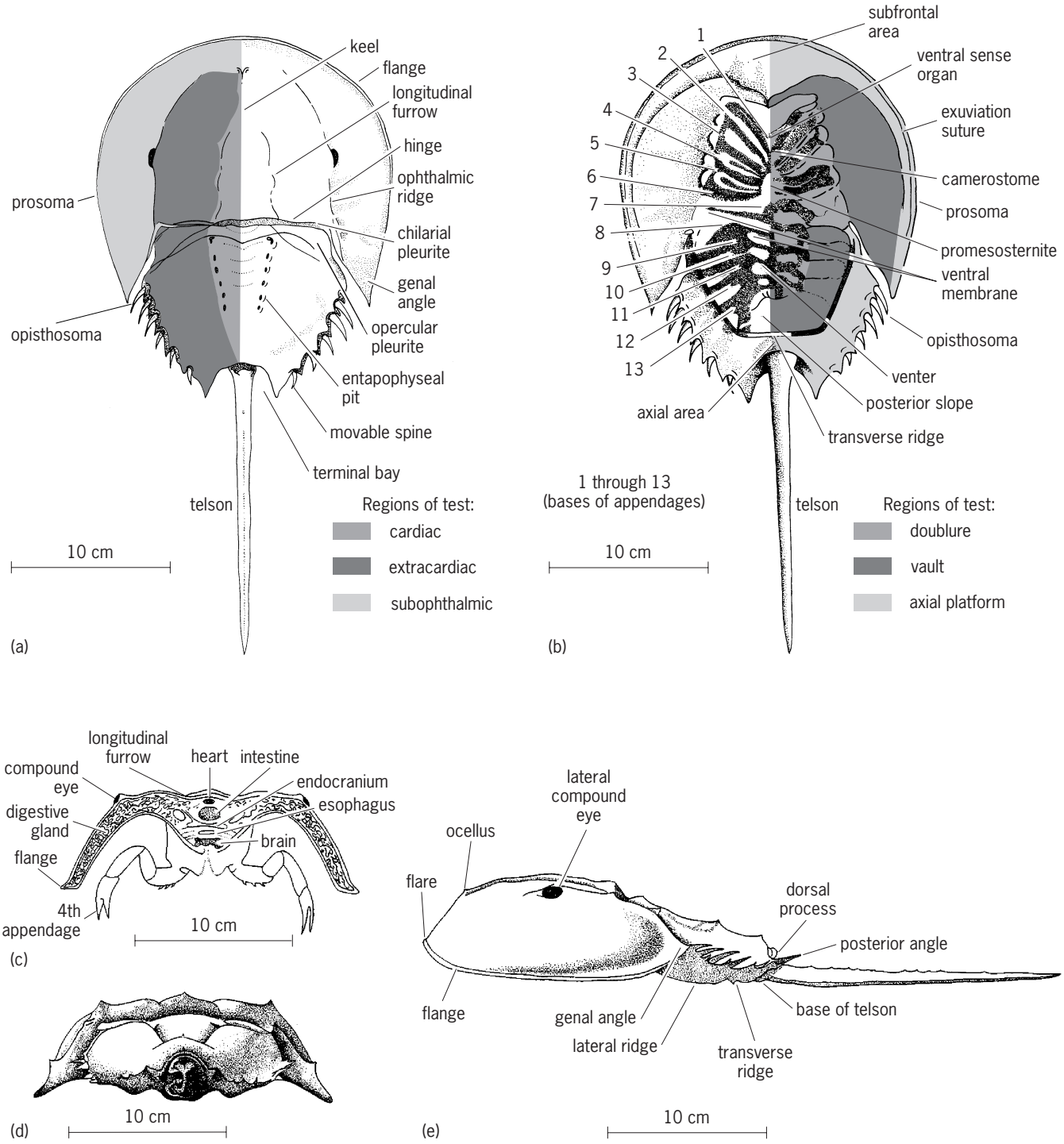


Fig. 6. Morphology of adult male *Limulus polyphemus*. (*a*) Dorsal view. (*b*) Ventral view, appendages not shown. (*c*) Cross section through prosoma showing internal anatomy. (*d*) Posterior view. (*e*) Lateral view.

tenfold. Just how the animals use these capabilities is unknown; perhaps they are used in sensing the time for spawning and for guiding the annual migration from deep water to the beaches where spawning takes place. The curved cone shape of the eye lenses has been incorporated into a solar energy collector. The major biomedical applications of limulines are the detection of bacterial endotoxin, using a lysate from the amebocytes, and the probing of cellular membrane constituents, using agglutinins (lectins) from the hemolymph. Limulines have some interesting medical uses. For example, they have been used to produce surgical sutures and wound dressings.

In some areas, as in New England waters and off New Jersey where shellfish resources are heavily harvested (the softshell clam, *Mya arenaria*, and the surf clam, *Spisula solidissima*, respectively), predation by *L. polyphemus* is a problem, and horseshoe crabs have been hunted and destroyed.

The taking of horseshoe crabs for various human uses and to prevent predation has seriously impacted

some *L. polyphemus* populations and, in conjunction with destruction of estuarine habitats, has raised concerns about the conservation of the species.

Carl N. Shuster, Jr.; Loren E. Babcock

Bibliography. L. I. Anderson and P. A. Selden, *Lethaia*, 30:19–31, 1997; L. E. Babcock, D. F. Merriam, and R. R. West, *Lethaia*, 33:129–141, 2000; P. S. Borkow, and L. E. Babcock, *The Sedimentary Record*, 1(3):4–7, 2003; D. C. Fisher, in N. Eldredge and S. M. Stanley (eds.), *Living Fossils*, 1984; J. H. Lochhead, in F. A. Brown (ed.), *Selected Invertebrate Types*, 1950; S. M. Manton, *The Arthropoda: Habits, Functional Morphology, and Evolution*, 1977; S. P. Parker (ed.), *Synopsis and Classification of Living Organisms*, 2 vols., 1982; C. N. Shuster, Jr., *The Circulatory System and Blood of the Horseshoe Crab, Limulus polyphemus L.: A Review*, 1978; C. N. Schuster, Jr., R. B. Barlow, and H. J. Brockmann (eds.), *The American Horseshoe Crab*, 2004; R. E. Snodgrass, *A Textbook of Arthropod Anatomy*, 1952; L. Størmer, in R. C. Moore (ed.), *Treatise on Invertebrate Paleontology*, pt. P2, 1955.

Xylem

The principal water-conducting tissue and the chief supporting system of higher plants. This tissue and the associated phloem constitute the vascular system of vascular plants. Xylem is composed of various kinds of cells, living or nonliving. The structure of these cells differs in their functions, but characteristically all have a rigid and enduring cell wall that is well preserved in fossils.

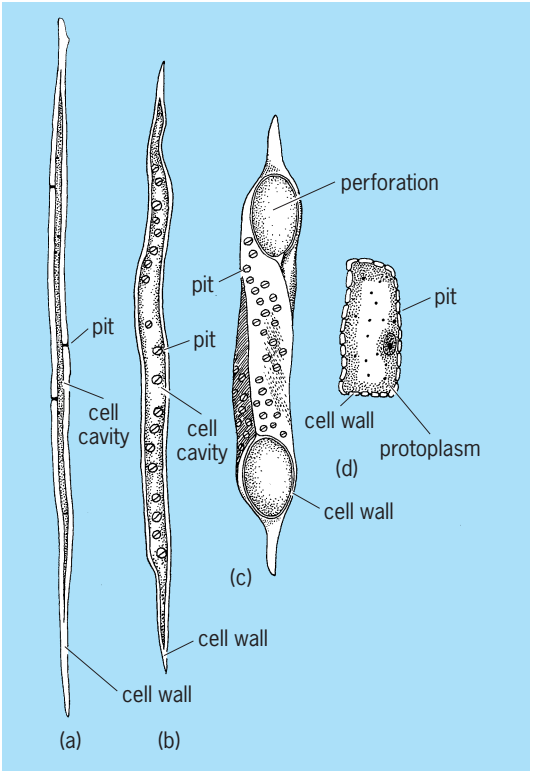


Fig. 1. Xylem cell types. (*a*) Wood (xylem) fiber. (*b*) Tracheid. (*c*) Vessel member. (*d*) Xylem parenchyma cell. All can vary widely in structure. (*After H. J. Fuller and O. Tippo, College Botany, rev. ed., Holt, 1954*)

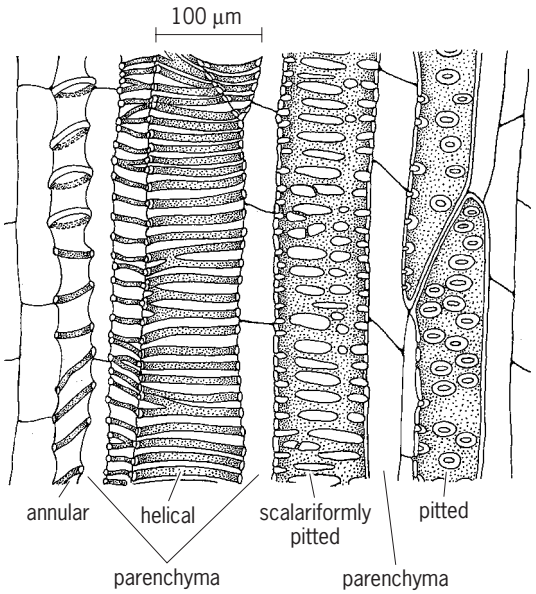


Fig. 2. Parts of tracheary elements and associated parenchyma from stem of *Aristolochia*, as seen in longitudinal section. The secondary walls are stippled. The earliest part of the xylem is at the left. (*After K. Esau, Plant Anatomy, 2d ed., John Wiley and Sons, 1965*)

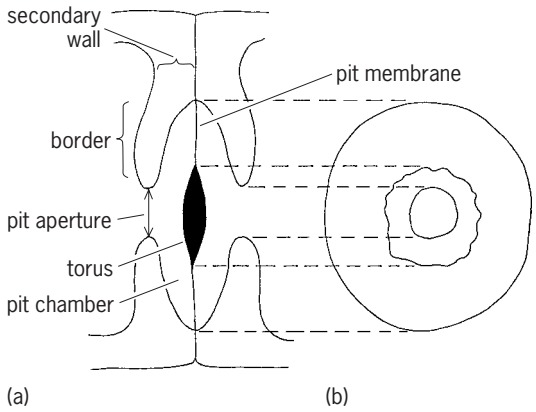


Fig. 3. Diagrams of bordered pit pairs of the genus *Pinus*. (*a*) Sectional view. (*b*) Face view. The torus is formed by thickening of primary walls. (*After K. Esau, Plant Anatomy, 2d ed., John Wiley and Sons, 1965*)

Kinds of xylem cells. In terms of their functions, the kinds of cells in xylem are those related principally to conduction and support, tracheids; to conduction, vessel members; to support, fibers; and to food storage, parenchyma (**Fig. 1**). Vessel members and tracheids are often called tracheary elements. The cells in each of the four categories vary widely in structure.

Tracheids. Tracheids are elongate, spindle-shaped cells, lacking protoplasm at maturity, and having secondary walls laid in various thicknesses and patterns over the primary wall. The secondary wall of tracheids and other conducting cells may consist of stretchable annular or helical (spiral) bands, or of a

nonextensible continuous expanse interrupted only at the bordered pits (**Fig. 2**).

Vessels. Vessel members differ from tracheids in having perforations, usually in the end walls common to superposed cells. A longitudinal series of vessel members constitutes a vessel. Although each varies markedly, tracheids are usually two to three times as long as vessel members, for example, 0.004–0.006 in. (1–1.5 mm) versus 0.02 in. (0.5 mm) in secondary wood. Vessels, however, may be from many centimeters to even meters in length.

In wood, pits are interruptions or recesses in the secondary wall, and are usually paired in common walls of contiguous cells. In bordered pits (**Fig. 3**) the secondary wall overarches the pit membrane (intercellular substance and two primary walls of contiguous cells); the overarch is absent in simple pits. The borders may be well developed or barely perceptible. The pit apertures (openings into the cell lumen) in bordered pits may be round to slitlike openings. Vertical series of transversely elongated pits are described as scalariform pitting. Round or oval pits may appear singly, or in horizontal (opposite pitting) or oblique (alternate pitting) series.

Perforations result from disintegration of the primary wall in the ends of vessel members. A perforation plate (**Fig. 4**) is the wall area bearing one or more perforations. The most primitive (scalariform) perforation plates are strongly oblique in position and have numerous transversely placed perforations (over 200). The most specialized have a single large (simple) perforation in a transverse wall.

Fibers. Fibers differ from tracheids in being more slender, in having simple pits and thicker walls in proportion to diameter, and in often retaining protoplasm. Fibers are thought to have evolved from tracheids, and intergrading forms between them are called fiber tracheids. Divisions of the protoplast of living fibers result in septate fibers.

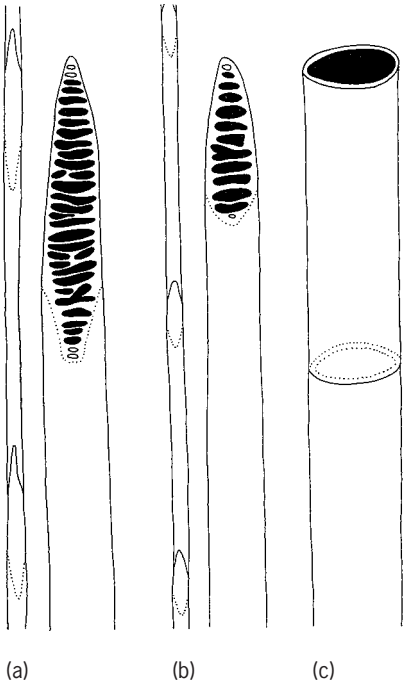


Fig. 4. Complete vessel members and parts of single-vessel members with perforation plates on end walls (perforations in black). (*a*) Long scalariform perforation plate. (*b*) Shorter scalariform. (*c*) Simple.

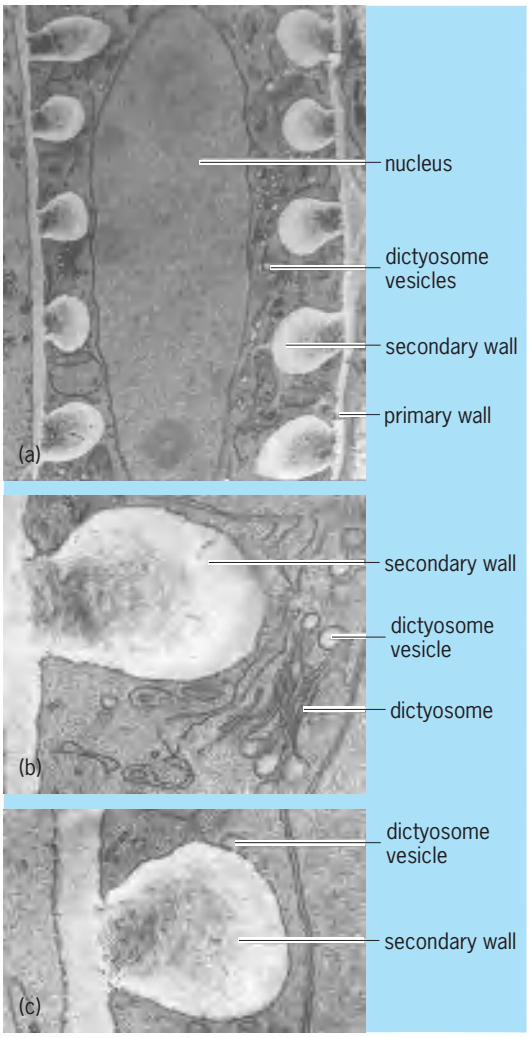


Fig. 5. Electron micrographs of longitudinal sections of vessel members with helical secondary wall. (*a*) Nucleus and numerous dictyosomes with vesicles. (*b*) One dictyosome near secondary wall. (*c*) Dictyosome vesicle attached to secondary wall after junction with plasmalemma coating the wall.

Parenchyma. The term parenchyma usually refers to the vertically arranged living cells of the xylem, although rays are also composed chiefly of parenchyma. Depending upon their origin, parenchyma cells may occur singly or as members of a strand. The cells may or may not have secondary walls, typically have simple pits, usually store starch, and may contain crystals, tannins, or other substances.

Development. Electron microscopy has shown that in the differentiating conducting cells the formation of secondary walls is associated with an intense activity of dictyosomes (Golgi bodies). These produce numerous vesicles which migrate toward the wall and fuse with the plasmalemma (plasma membrane) in

such a way that their contents appear to be released toward the secondary wall and their membranes become part of the plasmalemma (**Fig. 5**). The nucleus, all organelles, and the membrane systems degenerate as the conducting cell matures. If this cell is a vessel member, the perforation is formed before the protoplast dies.

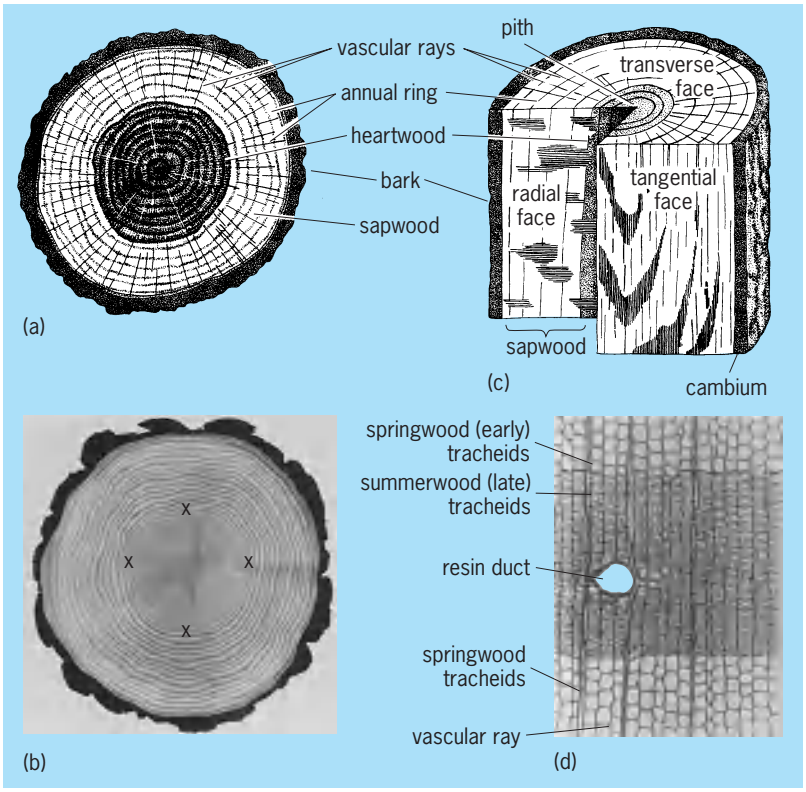


Fig. 6. Growth layers. (*a*) Cross section of oak branch. (*b*) Cross section of loblolly pine trunk showing effect of light on annual ring thickness. The annual rings inside the x's were formed while the young tree was densely shaded by older trees. The annual rings outside the x's were formed after the older trees had been removed and the young pine tree received abundant sunlight. (*c*) Three-dimensional appearance of a log. (*d*) Highly magnified part of transverse section of pinewood with parts of two annual rings. (*After H. J. Fuller and O. Tippo, College Botany, rev. ed., Holt, 1954*)

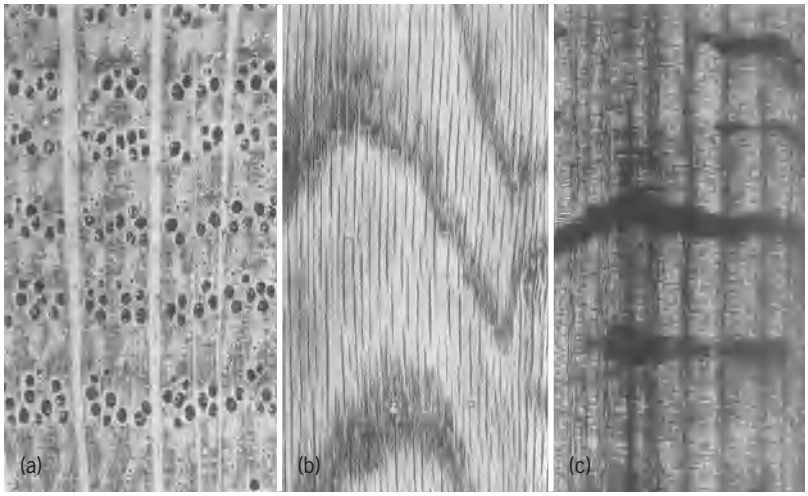


Fig. 7. Red oakwood (a ring-porous hardwood). The earlywood vessels are much larger than the latewood vessels. (*a*) Transverse section. (*b*) Tangential section. (*c*) Radial section. (*After H. J. Fuller and O. Tippo, College Botany, rev. ed., Holt, 1954*)

Primary and secondary xylem. Xylem tissues arise in later stages of embryo development of a given plant and are added to by differentiation of cells derived from the apical meristems of roots and stems. Growth and differentiation of tissues derived from the apical meristem provide the primary body of the plant, and the xylem tissues formed in it are called primary. Secondary xylem, when present, is produced by the vascular cambium. *See* LATERAL MERISTEM.

Primary. Protoxylem is that part of the primary xylem formed while the organ is still elongating and generally includes cells with extensible secondary wall thickening. These cells are frequently torn in vigorously growing plants. Metaxylem is formed after growth in length is nearly completed, and has cells mostly with nonextensible secondary walls.

Parenchyma cells, with thin or thick walls, occur in both protoxylem and metaxylem, and, depending upon the species, vessels, tracheids, and fibers all may be present.

Secondary. Secondary xylem is composed of two interpenetrating systems, the horizontal (ray) and vertical (axial). Wide variation occurs in secondary wood, mostly because of inherent differences among species and partly because of environmental influences. The differences in rate of evolutionary specialization of the different kinds of cells as seen in any one species make possible the positive identification of nearly all kinds of wood. Variations include differences in types of rays and their distribution; kinds, numbers, arrangement, and sizes of vessels, tracheids, and fibers; types of growth layers; character of sapwood and heartwood; kinds of knots; distribution of parenchyma; and other features, such as odor. Many of these variations bear on the strength and usefulness of wood.

Rays may be one to many cells in width as seen in transections of the organ in which they occur, may be composed of cells alike or unlike in form, are variable in length radially, and differ in height, for example, high in oak, low in maple. Evolutionary specialization has involved height and width of rays, even loss of rays, and changes in dimensions of the ray cells themselves. Activities of ray cells include the formation of tyloses—protrusions of parenchyma cells into the lumen of tracheary elements—and the secretion of variable quantities of gummy substances into neighboring cells.

A growth layer is the increment of secondary xylem produced in one growing season (**Fig. 6**). It is called an annual layer—or annual ring when seen in transection—if a single increment is formed each year. The boundaries between successive growth rings are identifiable largely because the xylem cells in the later part of one season (latewood or summerwood) are much smaller than the first ones produced in the next season's growth (earlywood or springwood). In ring-porous woods, the vessels are much larger in the earlywood than in the latewood of a given growth layer. The vessels are of approximately equal size throughout in diffuse-porous woods (**Fig. 7**).

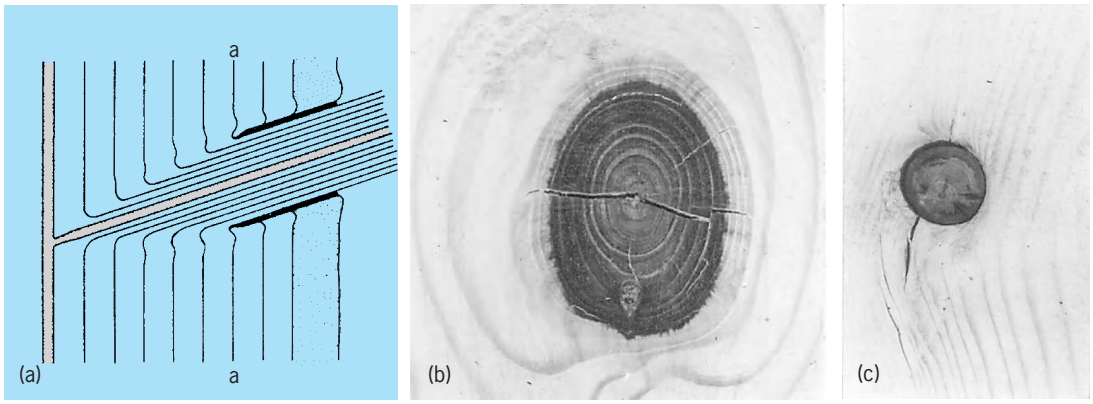


Fig. 8. Wood knots. (*a*) Diagram illustrating burial of base of a branch by secondary growth. The portion of the branch to the left of the line *a-a* was alive when buried, and its growth rings are continuous with those of the trunk. The portion to the right of *a-a* was dead when buried, and the growth rings of branch and trunk are not continuous. Pith finely stippled, bark coarsely stippled. (*b*) Hard pine board showing tight knot. (*c*) Hard pine board showing loose knot. (*After C. L. Wilson and W. E. Loomis, Botany, rev. ed., Dryden, 1957*)

Most knots are bases of twigs or branches buried, either alive (tight knot) or dead (loose knot), by cambial activity of the mother branch (**Fig. 8**). Their size depends largely upon the age at which they were buried. In lumber, round knots represent transections of the encased branch, and spike knots, transections of longitudinal sections.

Softwood and hardwood. In the trade, softwood is a name for xylem of gymnosperms (conifers) and hardwood for xylem of angiosperms. The terms do not refer to actual hardness of the wood. Woods of gymnosperms are generally composed only of tracheids, wood parenchyma, and small rays, but differ in detail (**Fig. 9**). Resin ducts are present in many softwoods. Woods of angiosperms show extreme variation in both vertical and horizontal systems, but with few exceptions have vessels (**Fig. 10**).

The grain, texture, and figure of wood reflect

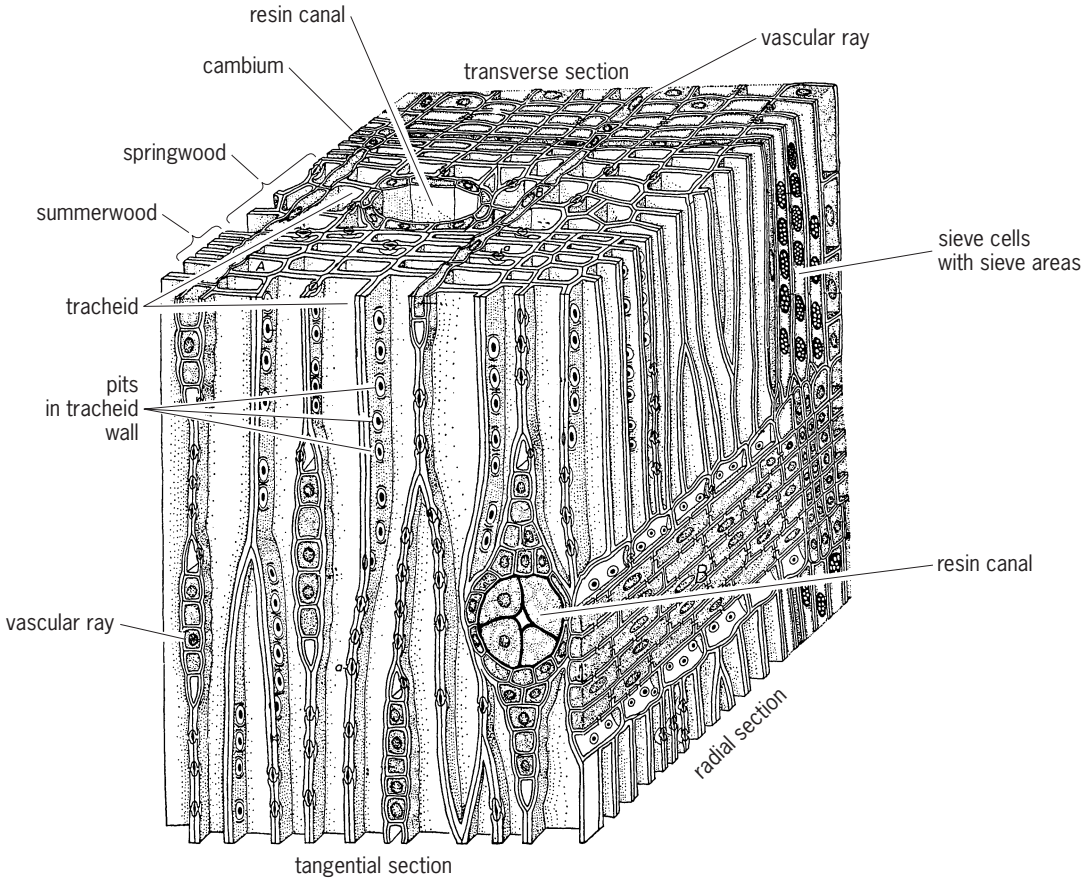


Fig. 9. Microscopic structure of pinewood (gymnosperm, three-dimensional). Gymnosperms are known as softwoods. (*After H. J. Fuller and O. Tippo, College Botany, rev. ed., Holt, 1954*)

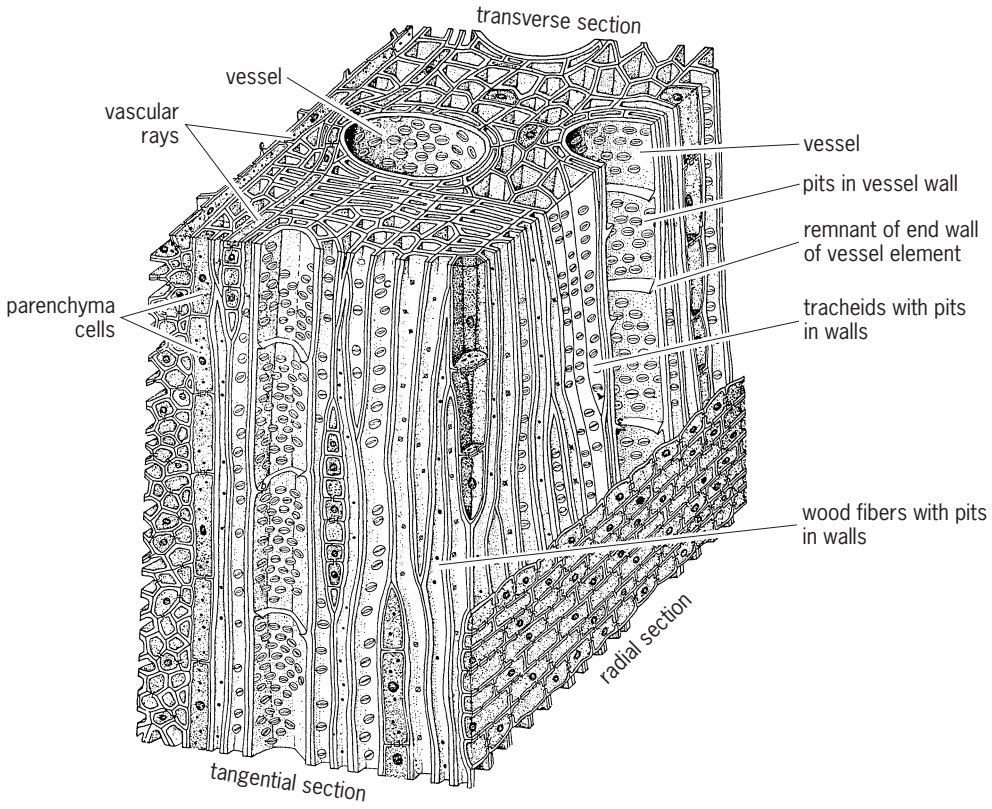


Fig. 10. Microscopic structure of oakwood (angiosperm, three-dimensional). Angiosperms are known as hardwoods. (*After H. J. Fuller and O. Tippo, College Botany, rev. ed., Holt, 1954*)

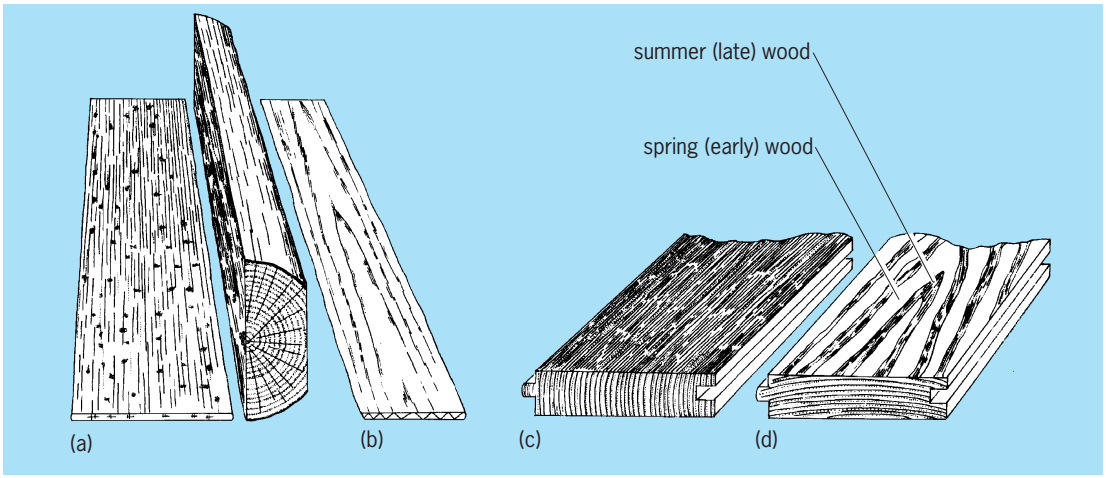


Fig. 11. Wood identification. (*a*) Quarter-sawed (radially sawed) and (*b*) plain-sawed (tangentially sawed) boards cut from a log (*after H. J. Fuller and O. Tippo, College Botany, rev. ed., Holt, 1954*). (*c*) "Edge-grain" flooring lumber quarter-sawed from fir trees. (*d*) Plain-sawed fir flooring, in which large uneven areas of springwood cause uneven wear and rapidly splintering surface (*after G. M. Smith et al., A Textbook of General Botany, 5th ed., Macmillan, 1953*).

gross differences in the arrangement, orientation, and kinds of cells in xylem and often are useful in identification (**Fig. 11**). These differences affect both beauty and strength of wood. Specific gravity also influences strength, as do the size, arrangement, and number of vessels, fibers, parenchyma cells, and rays. Heartwood, the older wood, is no stronger than the younger sapwood, although often it is more durable. Transforming sapwood into heartwood involves partial loss of water, death of living cells, production of various substances, often colored, and often the appearance of tyloses. *See* FOREST AND FORESTRY; PLANT TISSUE SYSTEMS; WOOD ANATOMY; WOOD PRODUCTS. Vernon I. Cheadle; Katherine Esau

Y-delta transformations

Relationships between electrically equivalent networks with three terminals, one being connected internally by a Y configuration and the other being connected internally by a Δ configuration. These relationships are also known as star-delta transformations. More complicated relationships can also be derived between a many-armed star and the equivalent many-sided mesh.

Before considering these networks, consider the simpler two-terminal network. If the path consists of a network, however complicated, of passive linear elements, the input impedance is the ratio of the transform or phasor of input voltage to the transform or phasor of entering current. It is usually not difficult to find the input impedance of any passive, linear, two-terminal network, commonly by series and parallel combination of impedances. The input impedance will in general be a function of frequency. At one particular frequency it is possible to design a single element to have the same impedance, and this element can be said to be equivalent to the given network at that frequency, in the sense that the terminal relations of voltage and current will be the same (at that frequency) for either. In certain particular networks the equivalence is valid at all frequencies.

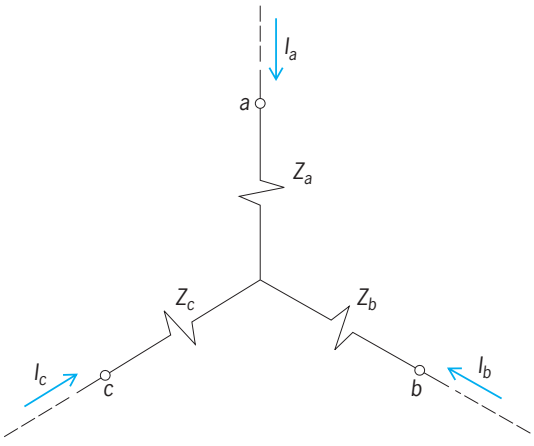


Fig. 1. Three-terminal star, or Y.

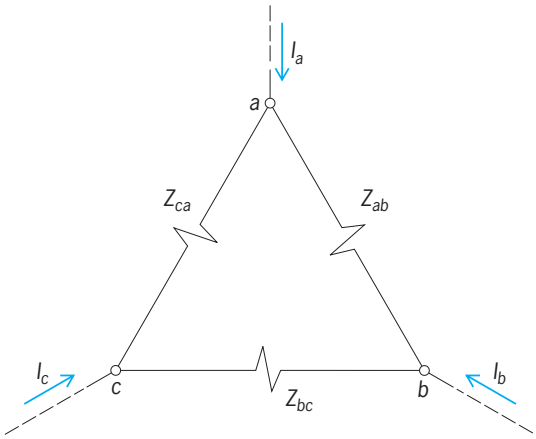


Fig. 2. Three-terminal mesh, or Δ.

If a network connects three terminals with one another, it is called a three-terminal network. The simplest configurations of three-terminal networks are the Y or star (**Fig. 1**) and the Δ or mesh (**Fig. 2**). [The star of three elements may also be called a T, and the mesh of three elements a π (**Fig. 3**).]

If a given three-terminal network, passive and linear, has a Y configuration, it is possible to design an equivalent Δ network that could be substituted for the Y without changing the relations of voltage and current at the network terminals, or elsewhere external to the network. Similarly, if a Δ network is given, an equivalent Y network can be found. Impedances of equivalent networks are usually functions of frequency, and realization of these impedances by use of physically possible elements is usually limited to a single frequency.

Derivation. If terminal-to-terminal voltages V_{ab}, V_{bc}, and V_{ca} are taken to be equal in the Y and Δ of Figs. 1 and 2, then the terminal currents I_a, I_b, and I_c are equal in the two networks if and only if the

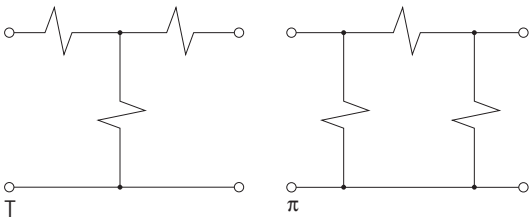


Fig. 3. Three-terminal star and mesh drawn as T and π.

impedance relations in Eqs. (1)–(3) are satisfied.

$$Z_a = \frac{Z_{ab}Z_{ca}}{Z_{ab} + Z_{bc} + Z_{ca}} \quad (1)$$

$$Z_b = \frac{Z_{bc}Z_{ab}}{Z_{ab} + Z_{bc} + Z_{ca}} \quad (2)$$

$$Z_c = \frac{Z_{ca}Z_{bc}}{Z_{ab} + Z_{bc} + Z_{ca}} \quad (3)$$

These are derived by expressing the three currents of each configuration in terms of the three voltages, equating, and solving simultaneously. Only one of these three equations is needed, for the subscripts a, b, and c are arbitrary (**Fig. 4**).

The same equivalence can be expressed in admittances instead of impedances by Eq. (4). Again, only

$$Y_a = \frac{Y_{ab}Y_{bc} + Y_{bc}Y_{ca} + Y_{ca}Y_{ab}}{Y_{bc}} \quad (4)$$

one equation is needed, for the subscripts are arbitrary.

To change Y to Δ, a simultaneous solution of Eqs. (1), (2), and (3) gives a set of relationships, Eqs. (5)–(7), to be used if the impedances or admittances of a Y are known and those of an equivalent Δ are wanted.

$$Z_{ab} = \frac{Z_aZ_b + Z_bZ_c + Z_cZ_a}{Z_c} \quad (5)$$

$$Y_{ab} = \frac{Y_aY_b}{Y_a + Y_b + Y_c} \quad (6)$$

$$Z_a = \frac{(20)(20)}{20 + 20 + j\omega 5} = \frac{400}{40 + j\omega 5} \text{ ohms} \quad (7)$$

For admittances, the equivalence is expressed by Eq. (6).

Example. The following example, though rather artificial, shows several interesting aspects of equivalence. Given the Δ (**Fig. 5**), with two sides purely resistive and the other side purely inductive, the problem is to find an equivalent Y. Given values are 20 ohms in each of two sides, and $L = 5$ henrys or $Z = j\omega 5$ ohms in the third side. From Eq. (1), relation (7) can be written. Similarly, from Eq. (2), relation (8) can be written. In this example $Z_c = Z_b$.

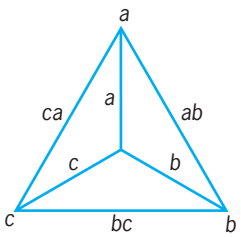


Fig. 4. Diagram to show the pattern of subscripts used in equations.

$$Z_b = \frac{(20)(j\omega 5)}{20 + 20 + j\omega 5} = \frac{j\omega 100}{40 + j\omega 5} \text{ ohms} \quad (8)$$

The impedances of the equivalent Y can be computed at any particular frequency; for example, at 60 cycles per second, or 60 Hz, $\omega = 377$. At low frequency, Z_a approaches 10 ohms of resistance, but for all finite frequencies it is somewhat less than 10 ohms and is more or less capacitive. Z_b and Z_c are equal because of symmetry in the original Δ network; at low frequency they have low inductive values, approaching zero, whereas at high frequencies each approaches 20 ohms of resistance.

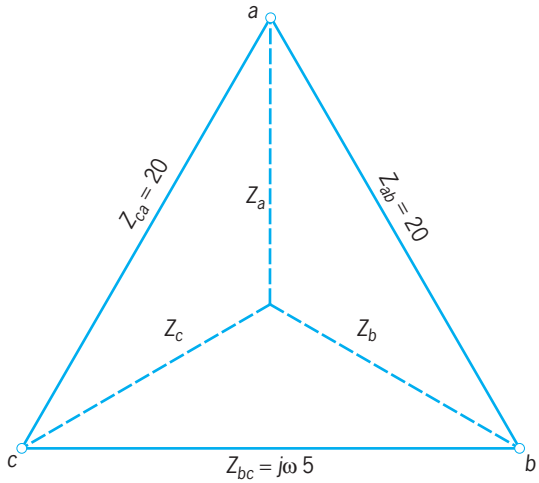


Fig. 5. Example of a Y that is equivalent to a given Δ.

At any particular frequency, each of the Y impedances can be physically realized as resistance in series with either a capacitance or an inductance, but physical realization is not simple over any range of frequencies.

Difficulty with physical realization, however, does not invalidate the mathematical equivalence, and the Y network that has here been determined can be substituted for the given Δ network for all analytical or computational purposes. *See* ALTERNATING-CURRENT CIRCUIT THEORY; NETWORK THEORY.

Hugh H. Skilling

Bibliography. R. DeCarlo and P.-M. Lin, *Linear Circuit Analysis*, 2d ed., 2001; J. D. Irwin and R. M. Nelms, *Basic Engineering Circuit Analysis*, 8th ed., 2005.

Yagi-Uda antenna

An antenna in which the gain of a single dipole element is enhanced by placing a reflector element behind the dipole (the driver) and one or more

director elements in front of it (see **illus.**). Invented in 1926 by H. Yagi, and S. Uda in Japan, the Yagi-Uda antenna is a long-time favorite of amateur radio enthusiasts. The gain is slightly increased by the reflector and further enhanced by the first director element. Additional director elements further increase the gain and improve the front-to-back ratio, up to a point of diminishing returns. This type of antenna has traditionally been used for local television reception. Its variants have found applications in the more modern communication systems at higher frequencies and smaller sizes, and have even been adapted to printed-circuit techniques in some applications. The same electromagnetic induction principle used in such linear elements can be applied to loop and disk elements as well with similar results. *See* DIRECTIVITY; ELECTROMAGNETIC INDUCTION; GAIN; PRINTED CIRCUIT.

Although the resonant half-wave dipole driver exhibits bandwidth limitation, the Yagi-Uda antenna can be considered as a traveling-wave antenna, and therefore various combinations of element lengths, separations, and other factors can achieve a range of phase velocities. The traveling-wave feature of the Yagi-Uda antenna can often provide a higher usable bandwidth than the equivalent resonant antenna design. The primary design objective is therefore to select the number of directors and the length and separation of the driver, the directors, and the reflector so as to optimize the antenna gain and improve the front-to-back ratio. This task has been rather tedious, but computer-aided techniques have been developed that can help provide an optimal design for a given gain and frequency. These techniques include method of moment (MOM) and finite difference time domain (FDTD).

The physical dimensions for the driver are normally about a half wavelength or less. The reflector is usually slightly larger than the driver and positioned about a quarter of a wavelength behind it. The directors are slightly shorter than the driver and positioned at distances of approximately a quarter of a wavelength in front. The design for one frequency can usually be scaled up or down for lower and higher frequencies. For comparison, a typical television antenna can be measured in feet (1 ft = 0.3 m), while the same antenna at 2.4 GHz, the frequency of the wireless local-area network, measures in inches (1 in. = 2.5 cm). The **table** shows some typical performance data for the traditional Yagi-Uda antenna. *See* LOCAL-AREA NETWORKS.

Typical performance characteristics of a Yagi-Uda antenna

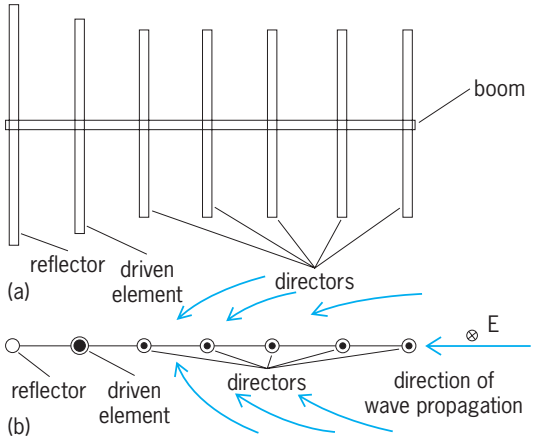
Number of elements	Gain dBi*	Beam width in E-plane†	Beam width in H-plane†
2	5	>66°	>84°
3	9	66°	84°
4	10	52°	60°
5	11	58°	66°

*Power gain, in decibels, in the direction of maximum radiation, compared to the radiation from an isotropic antenna in free space receiving the same amount of power.
†Full width of the primary lobe of the radiation pattern in the specified plane at half the maximum power.

Since these antennas can be made highly directive with good radiation efficiency, they have found new applications and new manufacturing techniques with miniaturization. They can be printed on microwave circuit substrates with high dielectric constants, which reduces their size even further. The parasitic electromagnetic coupling demonstrated in the Yagi-Uda antenna has been adapted to many new types of miniaturized antennas applicable to mobile communication devices in wide use, and will be used in future wireless Internet devices. *See* ANTENNA (ELECTROMAGNETISM); INTERNET; MICROWAVE SOLID-STATE DEVICES; MOBILE RADIO.

Jamal S. Izadian

Bibliography. R. C. Johnson, *Antenna Engineering Handbook*, 3d ed., McGraw Hill, New York, 1993; W. L. Stutzman and G. A. Theile, *Antenna Theory and Design*, 2d ed., Wiley, New York, 1997.



Yagi-Uda antenna. Typical radiation patterns are shown. (*a*) Side view with radiation pattern in the E plane (the plane containing the electric field). (*b*) Top view with radiation pattern in the H plane (the plane containing the magnetic field).

Yak

A member of the mammalian cattle family Bovidae in the order Artiodactyla. It is the least-known species of wild cattle alive today. Yaks (*Bos grunniens*) have been domesticated for at least several hundred years. They are powerful yet docile. Above 2000 m (6600 ft), they are the most useful domestic mammal. They are used for meat and milk production, as a mount, and as a beast of burden. The hair is used for making cloth and tents, the dung is used as fuel, and the tail serves as a flybrush.

Morphology. The yak, a relative of the bison, is a blackish-brown bovid with large black horns that grow out from the sides of the head and curve upward (see **illustration**). The horns, which are present in both sexes, are composed of bony cores attached to the frontal bones of the skull and hard

Adult male domestic yak (*Bos grunniens grunniens*). (*Copyright © Brent Huffman, 2000*)

sheaths of horny material. The massive body is covered by long guard hairs that almost reach the ground and a dense underfur of soft hair. The guard hairs form a shaggy fringe of coarse hair around the lower part of the body, the shoulders, flanks, and thighs. The shaggy muzzle may be white. The tail also consists of long hairs. The shoulders are high and humped, with a broad, drooping head. The limbs are short and sturdy with broad hooves and large dewclaws as an adaptation to mountainous environments. Males may attain a head and body length up to 325 cm (10.8 ft), a shoulder height up to 200 cm (6.6 ft), and a weight of 305–820 kg (670–1805 lb). Females are smaller and may weigh only one-third as much. Domesticated animals are considerably smaller than wild yaks and may be red, brown, or black in coloration.

Habitat and life history. Yaks are found in Tibet, China, India, Kashmir, and Nepal at elevations of 4000–6000 m (12,800–19,200 ft). They inhabit desolate steppe regions of alpine tundra and cold desert regions with swamps and moors. They are surefooted and are expert climbers. They spend a portion of the summer in the lower plains, but as the temperature rises they move to the higher plateaus. They swim and bathe in icy cold lakes and rivers. Yaks feed on grasses, herbs, mosses, and lichens. Since vegetation is very sparse, yaks must travel great distances from one grazing site to another. For most of the year, bulls live either a solitary existence or they live in small groups, whereas females and their young live in large herds of up to 200 animals. Breeding occurs in September and October. Following a gestation period of approximately 9 months, a single calf is born in June. Females produce a calf every other year because the nursing period lasts an entire year. Yaks reach their adult size in 6 to 8 years and live for about 25 years in the wild. Enemies include humans and packs of Tibetan wolves, which hunt solitary animals and calves that are separated from the herd.

The wild yak is classified as vulnerable by the World Conservation Union, is listed as endangered by the United States Department of the Interior, and is on Appendix 1 of CITES (Convention on International Trade in Endangered Species of Wild Fauna and Flora). Yaks are officially protected in China but are declining due to uncontrolled hunting. Organizing protective measures in the inaccessable mountain regions of Tibet is almost impossible; thus, there is little hope of saving the remaining wild yak populations. *See* ARTIODACTYLA; BISON.

Donald W. Linzey

Bibliography. *Grzimek's Encyclopedia of Mammals*, vol. 5, McGraw-Hill, 1990; D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Yaws

An infectious disease of humans caused by the spirochete *Treponema pertenue*. It is also known as frambesia and is largely confined to the tropics. Anecdotal information indicates that yaws still exists in parts of central Africa, South America, India, and Indonesia. Usually yaws is contracted in childhood by direct contact or from small flies feeding in succession on infected lesions and open wounds. No race or age possesses natural immunity.

Treponema pertenue is morphologically similar to *T. pallidum*. The general course of yaws resembles that of syphilis, but the individual yaws lesions tend to be larger and more persistent. The primary lesion or mother yaw occurs commonly on the lower extremities. Generalized lesions are often large and fungating. Destructive bone lesions are common, as are painful incapacitating lesions of the soles of the feet (crab yaws). Central nervous system and cardiovascular involvement occurs rarely if at all. Serological tests for syphilis are also positive in yaws, and the disease responds well to penicillin.

There are indications that while *T. pertunue* and *T. pallidum* are biologically and immunologically related, significant and persistent differences occur to justify distinct names for the two diseases. The relationship of yaws to other treponematoses, such as pinta, endemic syphilis, and bejel, has not been clearly determined. Mass survey and treatment campaigns directed primarily to the reduction of infectious cases have been conducted with success in many countries. The wide availability of penicillin since World War II has encouraged public health programs aimed at the reduction of the incidence of yaws, although much of the evidence remains anecdotal. Improvement in the quality of the environment in many tropical areas has also contributed to a reduction in incidence, although some foci of endemic yaws persist there. *See* SEROLOGY; SYPHILIS.

Thomas B. Turner

Bibliography. A. B. Koff and T. Rosen, Non-venereal treponematoses: Yaws, endemic syphilis and pinta, *J. Amer. Acad. Dermatol.*, 29(4):519–535, 1993; V. N. Sehgal et al., Yaws control/eradication, *Int. J. Dermatol.*, 33(1):16–20, 1994; T. B. Turner and D. W. Hollander, *Biology of the Treponematoses*, 1957.

Year

The period of the Earth's revolution around the Sun. This period can be measured in several ways.

Types of years. The most commonly used type of year, known as the tropical year, is the mean period for the Sun to go around the sky from one vernal equinox to the next. The tropical year varies slightly over time. Tropical year 2000 was approximately 365.242190 mean solar days, or 365 d 5 h 48 m 45.2 s. Seasons repeat with intervals of a tropical year.

The sidereal year is the average period of the Earth's revolution around the Sun until the Earth-Sun line returns to the same direction in space as measured by the star background. A sidereal year is approximately 365.25636 mean solar days long, or 365 d 6 h 9 m 10 s.

The anomalistic year is the average period of the Earth's return to perihelion in its orbit around the Sun. An anomalistic year is approximately 365.25964 days, or 365 d 6 h 13 m 53 s.

An eclipse year is the duration between alternate eclipse seasons, measured as the interval between the Moon passing alternate nodes. An eclipse year lasts only 346.62005 days, or 346 d 14 h 52 m 52 s.

A Gaussian year, defined as the period derived from Kepler's law for 1 astronomical unit, is 365.25690 days, or 365 d 6 h 9 m 56 s. *See* ASTRONOMICAL UNIT; KEPLER'S LAWS.

A Besselian year is based on the Besselian elements used to compute eclipses. *See* EARTH ROTATION AND ORBITAL MOTION; ECLIPSE.

Counting of years. The counting of years used through most of the world incorporates A.D. and B.C. for Anno Domini and Before Christ (or BCE, for Before Christian Era). This practice dates from the sixth-century scholar Dionysius Exiguus (Dennis the Small). The date of the birth of Jesus, however, is now believed to be several years before A.D. 1. The Venerable Bede, in the eighth century, defined 1 B.C. as the year before A.D. 1 and then counted backward from 1 B.C. This system, however, omits a year 0, so care must be taken in computing intervals that span dates from A.D. to B.C. The problem led to the invention in 1583 by Joseph Justus Scaliger of Julian Dates (J.D.), counting individual days from 4713 B.C. Julian Dates begin at noon, so half-integral Julian Dates correspond to midnight, when other calendar dates begin. January 0.5 in Julian year 2000 corresponds to J.D. 2451544.0. Besselian year 2000.0 corresponds to J.D. 2451544.533. Astronomical catalogs are now given for the year 2000, called J2000.0 to signify centuries of exactly 36525 days counted from January 0.5, 1900, corresponding to January 1.5, 2000, or J.D. 2451545.0.

The Julian year of exactly 365.25 days leads to a drift in the calendar, which led to the Gregorian calendar. This Gregorian calendar introduced the system of leap years that is still in use. *See* CALENDAR.

Extraterrestrial years. The term "year" is also used for the period of the orbit of other planets, both inside the solar system and around other stars. For example, a Jupiter year is 5.2 Earth years long. One of the planets around a distant star orbits in only 2 Earth days, so its year is only 2 Earth days long. *See* TIME.

Jay M. Pasachoff

Bibliography. P. K. Seidelmann (ed.), *Explanatory Supplement to the Astronomical Almanac*, 1992.

Yeast

A collective name for those fungi which possess, under normal conditions of growth, a vegetative body (thallus) consisting, at least in part, of simple, single cells. The cells making up the thallus occur in pairs, in groups of three, or in straight or branched chains consisting of as many as 12 or more cells. Vegetative reproduction is characterized by budding or fission. Sexual reproduction also occurs in yeast, and is differentiated from that of other fungi by sexual states that are not enclosed in a fruiting body. Yeasts are a phylogenetically diverse group of organisms that occur in two divisions of fungi (Ascomycotina and Basidiomycotina) and 100 genera. The 700 or more species that have been described possibly represent only 1% of the species in nature, so the majority of the yeasts have yet to be discovered. Yeast plays a large part in industrial fermentation processes such as the production of industrial enzymes and chemicals, food products, industrial ethanol, and malt beverage and wine; in diseases of humans, animals and plants; in food spoilage; and as a model of molecular genetics. *See* DISTILLED SPIRITS; MALT BEVERAGE; WINE.

Characteristics

The shape and size of the individual cells of some species vary slightly, but in other species the cell morphology is extremely heterogeneous. The shape of yeast cells may be spherical, globose, ellipsoidal, elongate to cylindrical with rounded ends, more or less rectangular, pear-shaped, apiculate or lemon-shaped, ogival or pointed at one end, or tetrahedral (**Fig. 1**). The diameter of a spherical cell may vary

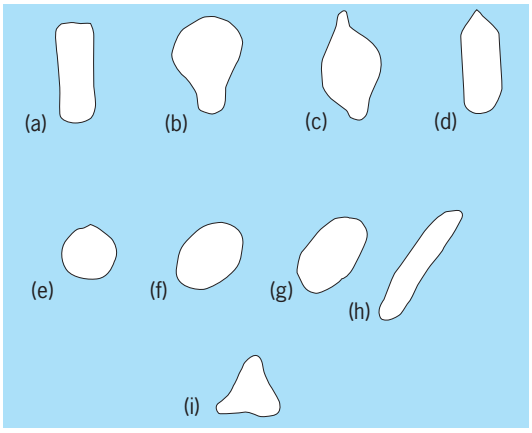


Fig. 1. Cell shapes of yeasts are extremely varied. (*a*) Rectangular. (*b*) Pear-shaped. (*c*) Apiculate. (*d*) Ogival. (*e*) Spherical. (*f*) Globose. (*g*) Ellipsoidal (oval). (*h*) Elongate. (*i*) Tetrahedral.

from 2 to 10 micrometers. The length of cylindrical cells is often 20–30 μm and, in some cases, even greater.

The asexual multiplication of yeast cells occurs by a budding process, by the formation of cross walls or fission, and sometimes by a combination of these two processes. Yeast buds are sometimes called blastospores or blastoconidia. When yeast reproduces by a fission mechanism, the resulting cells are termed arthrospores or arthroconidia. Some yeasts reproduce by cells known as ballistoconidia, which form on the parent cell at the tips of spear-shaped projections known as sterigmata. The ballistoconidia are forcefully ejected by a surface-tension catapult mechanism at an acceleration of 25,000 *g*, which is about 10,000 times the acceleration during the launch of the space shuttle. Whereas the shuttle consumes 50% of its weight in fuel in the first 2 minutes of flight, ballistoconidia discharge is fueled by carbohydrates that represent only 1% of the mass of the conidia.

In some genera of yeast, true septate mycelium is formed, for example, in some species of *Candida* and always in *Trichosporon*. In *Candida*, the mycelium, if present, remains intact; that is, it does not break up at the septa. In *Trichosporon*, the mycelium breaks apart into arthrospores which at first are rectangular but often develop rounded ends when mature (**Fig. 2***a*). Many yeasts produce a type of thallus which is termed a pseudomycelium. In a pseudomycelium the component cells are formed by elongation of buds and not by the formation of septa. Usually a group of characteristically arranged blastospores develops at the junctions of cells making up a pseudomycelium (Fig. 2*b*).

Yeasts are also characterized by certain macromorphological features. These are studied as slant cultures and as giant colonies. Slant cultures are made by inoculating the yeast as a thin line on the center of an agar slope in a medium which may be malt extract or synthetic. After several weeks of growth the culture may be described with respect to color (for example, cream, pink, yellow, brown, black, white, and intermediate shades); surface (smooth, wrinkled, warty, glossy, dull); texture (pasty, slimy or mucilaginous, tough); cross section (flat, raised, hemispherical, extent of spreading); and border (entire, hairy, irregular). Giant colonies are usually inoculated on malt gelatin, by a light, pinpoint inoculation. Macromorphological features usually develop more characteristically on gelatin than on agar media (**Fig. 3**). The description of the growth after about 3 weeks is similar to that used for the description of streak cultures.

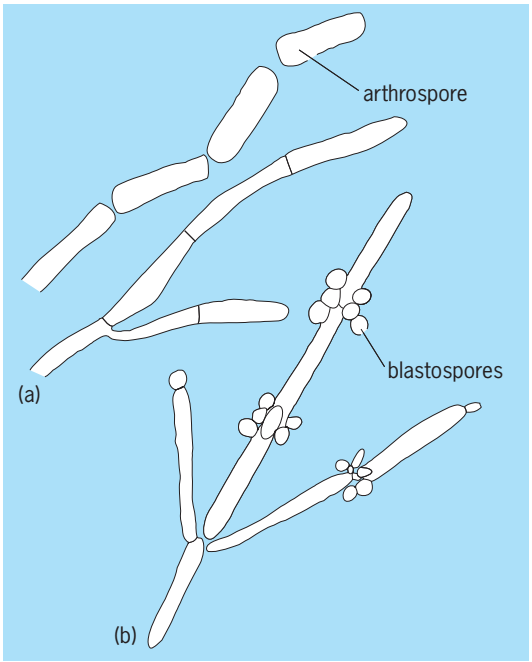


Fig. 2. Yeast mycelia. (*a*) True septate showing break to form arthrospore. (*b*) Pseudomycelium.

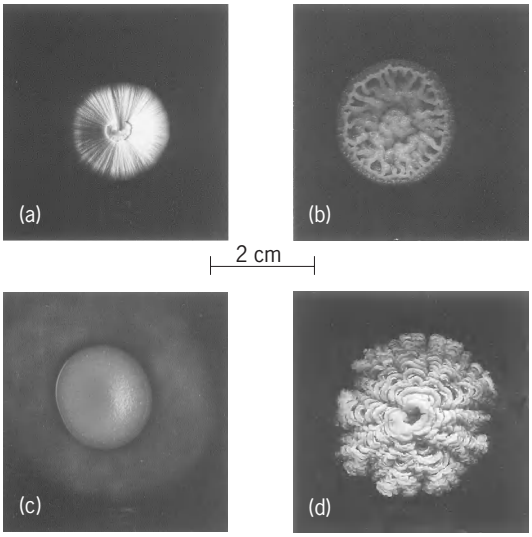


Fig. 3. Giant yeast colonies: (*a*) *Zygosaccharomyces rouxii*, (*b*) *Sporobolomyces pararoseus*, (*c*) *Rhodotorula glutinis*, and (*d*) *Saccharomyces cerevisiae*, all after 6 weeks of growth on wort gelatin at 68°F (20°C).

Cytology. Yeast cells are surrounded by a wall, which in the case of bakers' yeast and many other species consists of polysaccharides such as glucan and mannan, a small amount of chitin, protein, lipids, and minerals. The filamentous yeasts have a higher chitin content than the budding yeasts. Some species contain as yet unidentified components. With the electron microscope, bud scars can be observed in the walls of yeast. Successive buds, 20–40 from one bakers' yeast cell, are always formed at different places on the cell surface. A cell also contains a birth scar, which differs in appearance from a bud scar. Yeast cells are generally uninucleate. The cytoplasm may contain one or more vacuoles. In addition, the cytoplasm may contain lipid globules, volutin (polyphosphate) granules, mitochondria, and submicroscopic particles (microsomes). When yeast cells are treated with iodine, they usually stain deep brown, because of their high glycogen content.

Sexual reproduction. Yeasts are categorized into two groups, based on their methods of sexual reproduction: the ascomycetous (Division Ascomycotina)

and basidiomycetous (Division Basidiomycotina) yeasts. *See* ASCOMYCOTA; BASIDIOMYCOTA.

Ascospores. The sexual spores of the ascomycetous yeasts are termed ascospores, which are formed in simple structures, often a vegetative cell. Such asci are called naked asci because of the absence of an ascocarp, which is a more complex fruiting body found in the higher Ascomycetes. If the vegetative cells are diploid, a cell may transform directly into an ascus after the $2n$ nucleus undergoes a reduction or meiotic division. *See* MEIOSIS.

In the case of haploid yeasts, a conjugation or fusion between two cells normally occurs. The fusion of the cytoplasmic contents of the two cells is called plasmogamy. This is followed by fusion of the two haploid nuclei, or karyogamy. Meiosis then follows. Each of the haploid nuclei resulting from meiosis gathers cytoplasmic material around it and finally develops into a characteristically shaped ascospore. The spores in such a process are formed by "free cell formation," in contrast to spore delineation by cleavage of the cytoplasm. The ascospores may be spherical, oval, kidney- or crescent-shaped, hat-shaped, helmet-shaped, hemispherical, needle-shaped, walnut-shaped, or saturn-shaped (**Fig. 4**).

The number of ascospores per ascus varies with the species. Some contain a single spore, others one or two, two to four, or four to eight. In rare cases, more than eight spores are formed. In some yeasts, the asci rupture soon after maturation, and the spores are liberated into the medium. In others, the spores remain in the ascus until they germinate. The heat resistance of ascospores is only slightly greater than that of the vegetative cells. The number and shape of the ascospores are important characters in generic descriptions. The species of most genera have only one type of ascospore (such as the smooth spheroidal type in *Saccharomyces*), but sometimes more than one shape is found among the species of a genus, for example, in *Kluyveromyces*, *Hanseniaspora*, and *Saccharomycopsis*.

The ease with which yeasts form ascospores varies enormously with the species. In some, an abundance of ascospores is formed on nearly all media of growth. In others, special sporulation media must be used, and often the percentage of asci is small. Another complication is that certain yeasts lose the ability to sporulate upon prolonged cultivation in pure culture.

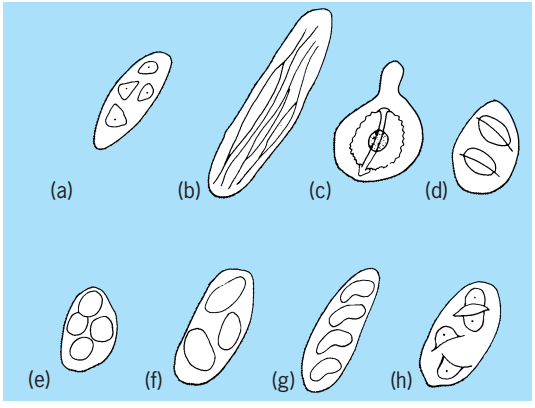


Fig. 4. Shapes of ascospores of various yeasts. (*a*) Helmet. (*b*) Needle. (*c*) Walnut. (*d*) Saturn. (*e*) Spherical. (*f*) Oval. (*g*) Kidney or crescent. (*h*) Hat.

Anamorphic and teleomorphic yeasts. Certain yeasts have been shown to be heterothallic; that is, sporulation occurs when strains of opposite mating type (usually indicated by a and α) are mixed on sporulation media. However, some strains may be homothallic (self-fertile), and reduction division and karyogamy take place during formation of the sexual spore. Yeasts that produce sporogenous cells represent the teleomorphic form of the life cycle. In cases, in which sexual cycles are unknown, the yeast represents the asexual or anamorphic form. A species of yeast may be originally discovered in the anamorphic form and named accordingly; subsequently, the sexual state may be found and a name applied to represent the teleomorph. Consequently, the anamorphic and teleomorphic names will differ. For example, the human pathogen *Cryptococcus neoformans* was originally found in the anamorphic form; several decades later the sexual cycle was discovered, and the teleomorphic form *Filobasidiella neoformans* was described. However, not all teleomorphic species have anamorphic counterparts. For example, the teleomorphic beer and bread yeast *Saccharomyces cerevisiae* was described in 1883, and anamorphic names for it are generally not recognized.

Among the majority of the yeast species, the teleomorphic forms are unknown; therefore, a separate classification scheme was created for the anamorphic yeasts, and orders such as Cryptococcales and Sporobolomycetales were erected. However, with the advent of molecular sequence analysis and the phylogenetic interpretation of the data, the relationships between anamorphic and teleomorphic species are becoming established. Consequently, the anamorphic species are being classified with the related teleomorphs, and the use of anamorphic orders is being eliminated.

Basidiospores and teliospores. These sexual spores are produced in the three classes of basidiomycetous yeasts: Urediniomycetes, Hymenomycetes, and Ustilaginomycetes. Sexual reproduction and life cycle in these yeasts is typical of other basidiomycetes in that it can include both unifactorial (bipolar) and bifactorial (tetrapolar) mating systems. The Urediniomycetes include species of widespread occurrence in the teleomorphic genera (*Leucosporidium*, *Rhodosporidium*, and *Sporidiobolus*) and the anamorphic genera (*Bensingtonia*, *Rhodotorula*, and *Sporobolomyces*). Reproduction in this class can include the formation of teliospores (**Fig. 5**). Conjugation between strains of opposite mating types (for example A1B1 × A2B2) results in dikaryotic mycelia with clamp connections. Karyogamy occurs in the heavy-walled teliospore, which germinates to produce a two- to four-celled septate metabasidium. Meiosis takes place during germination, and the resulting basidiospores are haploid.

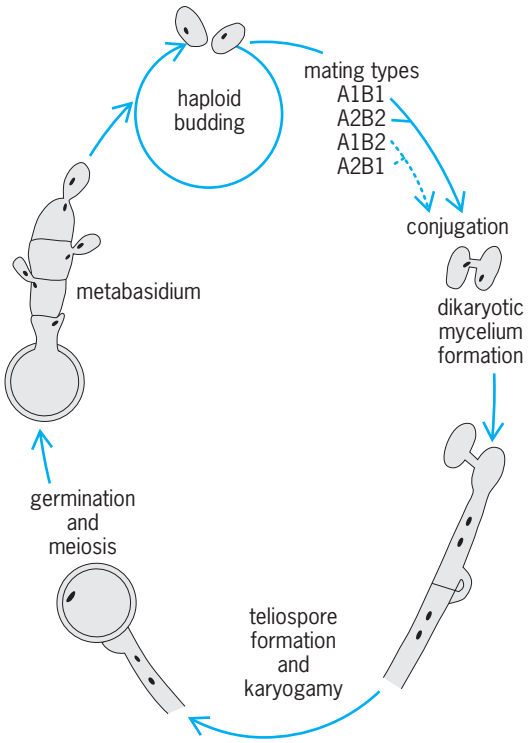


Fig. 5. Heterothallic (bifactorial) life cycle of the basidiomycetous yeast *Leucosporidium scottii*.

The Hymenomycetes include genera such as *Cryptococcus*, *Trichosporon*, *Bullera*, *Filobasidiella*, and *Filobasidium*. Some of these species are related to the jelly fungi, and others are well-known human pathogens. Sexual cycles among these yeasts are similar to the sexual cycles of the Urediniomycetes, except that most species do not produce teliospores, and the metabasidium is often an elongate single cell with an enlarged circular terminal head that bears the basidiospores.

The Ustilaginomycetes include the smut fungi. Yeasts in this class include species of *Malassezia*, *Pseudozyma*, and *Rhodotorula*. Sexual cycles among these yeasts are not well known.

Physiology. Some yeasts have the ability to carry out an alcoholic fermentation. Other yeasts lack this property. In addition to the fermentative type of metabolism, fermentative yeasts as a rule have a respiratory type of metabolism, whereas nonfermentative yeasts have only a respiratory, or oxidative, metabolism. Fermentation may be represented by reaction (1), respiration of glucose by reaction (2).

$$C_6H_{12}O_6 \rightarrow 2C_2H_5OH + 2CO_2 \quad (1)$$

$$C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O \quad (2)$$

Both reactions produce energy, with respiration producing by far the most, which is used in part for synthetic reactions, such as assimilation and growth. Part is lost as heat. In addition, small or sometimes large amounts of by-products are formed, including organic acids, esters, aldehydes, glycerol, and higher alcohols. When a fermenting yeast culture is aerated, fermentation is suppressed and respiration increases. This phenomenon is called the Pasteur effect.

Fermentation. Yeasts are best known for their ability to ferment mono- or disaccharides. In the case of disaccharides, trisaccharides, and polysaccharides, hydrolysis by hydrolytic enzymes, or hydrolases, must precede fermentation. There is evidence that some hydrolytic enzymes are located in the peripheral layer of the cell. Such hydrolases may be secreted into the medium as extracellular enzymes. Species of yeast are often differentiated on the basis of the various hydrolases which they possess. *See* CARBOHYDRATE.

Some simple rules of fermentation have been formulated. If a yeast cannot ferment glucose, it cannot ferment any other sugar. If a yeast can ferment glucose, it can also ferment fructose and mannose (although often at different rates), but not always galactose. If a yeast ferments maltose, it usually cannot ferment lactose and vice versa. *See* FERMENTATION.

Another mechanism of alcoholic fermentation has been discovered with the pentose sugar D-xylose as a substrate. Pentose sugars used to be thought of as nonfermentable, but it is now known that some yeast species (for example, *Pachysolen tannophilus* and *Candida tropicalis*) can convert D-xylose aerobically by a complex process, involving isomerization and phosphorylation, to fructose-6-phosphate which is fermentable by the classical pathway. The alcohol yields, however, are considerably lower than those obtained from glucose by conventional yeasts (*Saccharomyces cerevisiae*).

Respiration and assimilation. Yeasts differ greatly in their ability to respire and to assimilate organic substrates. Depending on the species, yeasts can utilize such compounds as pentoses (D-xylose, D-arabinose, L-arabinose, D-ribose), methyl pentoses (L-rhamnose), sugar alcohols (mannitol, sorbitol, dulcitol, erythritol, adonitol), organic acids (lactic, acetic, succinic, citric, gluconic, 2-ketogluconic, 5-ketogluconic, and many other acids), polysaccharides (soluble starch, inulin), *i*-inositol, and even such compounds as phenol and related ones.

The source of nitrogen may be organic or inorganic. All yeasts, with rare exceptions, can use ammonium ion for the synthesis of proteins. Others can utilize nitrate, and still others can use nitrite but not nitrate. Yeasts also can absorb intact amino acids from the medium. There is considerable variation in the ability of yeasts to deaminate individual amino acids. Lysine, as a single nitrogen source, is used by few yeasts. Glutamic acid is utilized by nearly all yeasts. Sulfur is ordinarily supplied as sulfate, although some yeasts grow better when the sulfur is supplied as cysteine or methionine. The following mineral salts and concentrations are required by yeasts in a medium: monobasic potassium phosphate, KH_2PO_4, 0.1%; magnesium sulfate, $MgSO_4$, 0.05%; sodium chloride, NaCl, 0.01%; and calcium chloride, $CaCl_2$, 0.01%. In addition, the following trace elements should be added to a synthetic medium: boron, copper, zinc, iron, manganese, molybdenum, and iodine.

Special synthetic products are carotenoid pigments, produced by species of *Rhodotorula* and *Sporobolomyces* (β-carotene is responsible for a yellow color; torulin and torularhodin are responsible for pink colors). Some species of the genus *Cryptococcus* produce a starchlike polysaccharide in acidic media. In addition, species of *Cryptococcus*, *Rhodotorula*, and some other yeasts produce capsules, consisting of various polysaccharides.

Ecology. Yeasts are ubiquitous in nature. They exist on plants and animals; in waters, sediments, and soils; and in terrestrial, aquatic, and marine habitats. Yeasts require oxygen for growth and reproduction; therefore they do not inhabit anaerobic environments such as anoxic sediments. Many species have highly specific habitats, whereas others are found on a variety of substrates in nature.

Soil. Soil may be considered as a reservoir of yeasts. Active growth in soil occurs only under favorable conditions, such as those found in fruit orchards and meadowlands. However, small numbers of a great variety of yeasts are found in many types of soil. Yeasts belonging to the genera *Lipomyces* and *Schwanniomyces* have been isolated only from soil.

Trees. Other yeasts have their natural habitat in the exudates or slime fluxes of trees. The yeasts found in exudates of coniferous trees generally differ from those in exudates of deciduous trees. Species of *Nadsonia* have been found only in tree exudates; certain species of *Hansenula*, *Saccharomycodes*, *Pichia*, *Saccharomyces*, and *Endomyces* are specifically associated with certain trees.

Insects and mammals. Since insects, such as *Drosophila*, frequently use sap exudates for breeding and sometimes for feeding purposes, yeasts can be found in appreciable numbers in the intestinal tract of such insects. Bark beetles, which live in the cambium layer of coniferous as well as deciduous trees, and the wood-boring ambrosia beetles also have been shown to have a specific yeast flora associated with them. Many species of yeast have been found in nature only in association with certain insects. Yeasts are rapidly digested by insects, and in this way they form an important part of their diet. The nectar of flowers is another habitat for yeasts. Those flowers, which are much frequented by bees, are especially likely to contain yeast. The intestinal tract of warm-blooded animals is also a habitat for yeasts. In some cases the yeast is so dependent on its host that it has lost the ability to grow at room temperature and in the common media used for propagation of yeasts. An example is the monospecific genus *Cyniclomyces*, which has been found only in the intestinal tract of rabbits. *Cyniclomyces guttulatus* grows only at about 99°F (37°C) and in very complex media. Interestingly, in the presence of about 15% free carbon dioxide the minimum temperature for growth is lowered to 86°F (30°C).

Pathogens. Whereas most yeasts associated with warm-blooded animals are nonpathogenic, a number of at least potentially pathogenic yeasts are known. With the expanding use of antibiotics, the natural balance of microbes in the gastrointestinal tract of mammals can be disturbed, and yeast infections are on the increase. The most common locations of infections by yeasts are the skin, especially the mucosa, and the respiratory tract, but occasionally systemic infections occur. *See* MEDICAL MYCOLOGY.

Plant pathogenic yeasts are limited to a few species. These are *Nematospora coryli*, a yeast with needle-shaped ascospores, and a yeastlike fungus, *Ashbya gossypii*, producing similar ascospores. The diseases which result from infection by *N. coryli* are yeast spot in lima beans and soybeans; stigmatomycosis of the fruits of tomatoes, citrus fruits, and others; and discoloration of cotton bolls. The cotton plant is also the principal host for *A. gossypii*. The yeast is transmitted by species of a number of genera of insects with piercing-sucking mouthparts as found in the bug group. Lastly, aquatic species of *Metschnikowia* are known as pathogens of the fresh-water crustacean *Daphnia magna* and the brine shrimp *Artemia salina*. *Metschnikowia bicuspidata* is the principal species that has caused economic losses in the brine shrimp industry. Infection and virulence may be related to the forceful ejection of the needle-shaped ascospores that pierce the intestinal wall of the host. Unidentified species of *Metschnikowia* also have been implicated in diseases of copepods during certain seasons of the year. *See* PLANT PATHOLOGY.

Food spoilage. Yeasts are important as spoilage organisms of a great variety of foods. Specific types occur, depending primarily on the composition of the food. Sugar-tolerant, or osmophilic, yeasts occur on dried fruits, syrups, honey, and the like. Examples are *Zygosaccharomyces rouxii*, *Schizosaccharomyces octosporus*, and the yeastlike fungus *Eremascus albus*. Species of *Debaryomyces* are highly salt-tolerant and occur in brines and on processed meats, such as bacon, hams, and sausage. Dairy products may be spoiled by lactose-fermenting yeasts. Some species of *Torulopsis* and of *Zygosaccharomyces* (haploid forms of *Saccharomyces*) may spoil salad dressing and similar products because of their great tolerance to vinegar. *Saccharomycopsis fibuligera* forms starch-splitting enzymes and can cause spoilage of cereal products and grains. *See* FOOD MICROBIOLOGY.

Jack W. Fell; Herman J. Phaff

Industrial Yeast

Yeasts are the most exploited group of industrial microorganisms. *Saccharomyces cerevisiae* (baker's or brewer's yeast) has been used for centuries in the production of fermented beverages and baked foods. In recent years, the industrial importance of *S. cerevisiae*, together with non-*Saccharomyces* yeasts, has extended beyond traditional food fermentation applications into the production of a wide range of important industrial commodities. **Table 1** summarizes the applications of several yeast species in the pharmaceutical, chemical, food, and environment sectors of biotechnology. There is great potential to further exploit the industrial power of yeasts,

TABLE 1. Some industrially important yeasts

Yeasts	Uses
Candida spp (for example, *C. utilis, C. maltosa, C. shehatae*)	Production of single-cell protein, vitamins, and organic acids.
Hansenula polymorpha	Expression of foreign genes for production of pharmaceutical proteins.
Kluyveromyces marxianus	Rich source of food-grade enzymes (lactase, pectinase, protease); ethanol production (from cheese whey fermentation).
Phaffia rhodozyma	Production of additive colorant (red-pink astaxanthin pigments) for farmed salmon feed.
Pichia pastoris	Production of single-cell protein and therapeutic proteins (from genetically manipulated strains).
Saccharomyces cerevisiae	Production of bread, alcoholic beverages, yeast extracts for food flavorings, fuel alcohol (bioethanol), glycerol, animal feed/growth factor, and biopharmaceutical protiens (through recombinant DNA technology).
Schizosaccharomyces pombe	Millet beer and rum production; wine deacidification; in vitro toxicity testing of mutagenic chemicals; potential in cloning technology.
Schwanniomyces occidentalis	Conversion of starch and inulin (a fructose polymer) into ethanol.
Yarrowia lipolytica	Production of protein and lipid material and industrial enzymes (lipases); it citric acid fermentation.
Zygosaccharomyces rouxii	Food-spoilage (being osmotolerant); production of Japanese soy sauce and miso.

particularly in the health care sector where genetically manipulated yeasts can be used to produce therapeutic proteins for combating human disease.

Fermented beverages. The production of ethyl alcohol (ethanol) by yeast fermentation is as old as human history. Yeast-derived ethanol now represents the premier global biotechnological commodity on both a quantitative and economic basis. Current worldwide production stands at around 30 billion liters per year and includes both potable alcohol (beverages such as beer, wine, cider, sake, whiskey, gin, rum, cognac, vodka, and tequila) and nonpotable alcohol (for fuel and industrial use, often referred to as bioethanol). In both cases, production relies on the ability of yeast cells (principally *S. cerevisiae*) to readily metabolize sugars into ethanol and carbon dioxide under anaerobic conditions using a multienzyme-catalyzed process known as fermentation.

The cereal-based fermented beverages include beer and whiskey (derived from barley) and sake (derived from rice). These products require fermentable sugars to be released from the cereal plant starch following the action of amylase enzymes, which degrade the starch (a polymer of glucose which is not fermented by yeast) into maltose, glucose, and other sugars which are fermented. For beer and whiskey production, this is achieved by the malting process (in which the barley grains are partially germinated to allow amylases to be synthesized) and the mashing process (in which crushed malt infused with warm water allows the amylases to break down the starch). In sake production, release of fermentable sugars from rice is achieved by allowing amylases from fungi such as *Aspergillus oryzae* to facilitate starch hydrolysis. For fruit-based beverages such as wine and cider, fermentable sugars (glucose and fructose) are released by simply crushing the fruit before fermentation proceeds. Yeasts responsible for wine production may be associated with the natural microflora of the grape skin surface (as in traditional winemaking) or specially selected starter cultures of *S. cerevisiae* (as in modern, large-scale winemaking). *See* ALCOHOL FUEL; ETHYL ALCOHOL; FOOD FERMENTATION.

Technological improvements in the production of fermented beverages include both genetic and physiological improvement of yeast strains as well as advances in fermenter design (for example, use of immobilized yeast cells). Genetic improvement of brewing and winemaking yeasts is possible using recombinant deoxyribonucleic acid (DNA) technology, and some strains have approval for food use. However, such yeasts are not currently being used due to adverse public reaction to "genetically manipulated" foods. *See* GENETIC ENGINEERING.

Baker's yeast. Yeast fermentation of sugars produces carbon dioxide in addition to ethanol. Carbon dioxide from yeast is exploited in the leavening of cereal doughs in the breadmaking process. In order to meet the yeast demands of the baking industry, specially selected strains of *S. cerevisiae* are propagated on a large scale. In fact, baker's yeast propagation represents the largest global bulk production of a single-celled microorganism, with several million tons produced annually. Prior to the development of separate baker's yeast industries around a hundred years ago, bakers used to rely on the availability of excess yeast (barm) from breweries. However, yeast from these sources is not suited for modern bakeries, and today baker's yeast is produced in large-scale multistage processes using molasses (sugarcane and sugarbeet refining residues) as a feedstock.

Molasses is a rich source of readily utilized sucrose, vitamins, and minerals for yeast growth. Other nutrients such as nitrogen (in the form of ammonia) and phosphorus (in the form of dibasic sodium phosphate) are added to the molasses prior to dilution, clarification, and flash sterilization (for example, 136°C for 15–30 seconds). Baker's yeast is propagated under fully aerobic conditions (at around 30°C and pH 4), and the rate of molasses feeding into the yeast propagators is carefully controlled to encourage respiratory metabolism and prevent the cells from fermenting and producing ethanol. In this

way, energy generation in the form of adenosine triphosphate (ATP) from respiration is maximized, and the production of yeast biomass is enhanced. At the end of the molasses feeding regime (referred to as fed-batch cultivation), aeration in the yeast propagators continues in order to condition or ripen the cells. This results in an increase in the levels of trehalose (a glucose disaccharide) within the yeast cells, which improves their storage stability and general stress tolerance. Yeast biomass is finally harvested by centrifugation and dried by rotary-drum vacuum filtration (to around 70% moisture) prior to packaging as fresh-pressed baker's yeast. Active-dried baker's yeast (at around 8% moisture) is prepared by spray-drying prior to packaging. *See* ADENOSINE TRIPHOSPHATE (ATP); CARBOHYDRATE METABOLISM.

Other applications of yeast biomass. In addition to their direct food use in the baking industry, yeasts may be specially grown to provide sources of protein and vitamins for human and animal nutrition. Several yeast species have been considered over the years for the production of single-cell protein, including *Candida utilis*, *C. tropicalis*, *Kluyveromyces marxianus*, *K. lactis*, and *S. cerevisiae*. These yeasts have the ability to grow on cheap, readily available carbon sources, such as straight-chain alkanes, corn steep liquor, cheese whey, sulfite waste liquor, and molasses, to produce high-quality protein. Although the use of yeast single-cell protein for direct human consumption has been discontinued by several companies, yeasts grown for use as animal feed and as a livestock growth factor are widespread. The use of live cultures of *S. cerevisiae* in ruminant animal feed has been shown to stabilize the rumen bacteria and improve nutrient availability to increase animal growth (and milk yields in the case of cattle). In human dietary supplements, yeast provides a source of B-complex vitamins, proteins, essential amino acids, and minerals. *See* AMINO ACIDS; PROTEIN; VITAMIN.

Some foods benefit directly from the presence of yeast extracts (whole yeast cells extracted with acids, enzymes, or salt) by imparting characteristic savory flavors and aromas. These compounds derive from the hydrolysis of yeast protein and ribonucleic acid (RNA) to yield taste-enhancing chemicals (for example, glutamic acid, peptides, and ribonucleotides). Some yeast extracts are prepared from excess brewer's yeast, but most are produced from strains of *S. cerevisiae* specially propagated for food applications. Yeast extracts are used in processed convenience foods to impart a meaty flavor (for example, dried soups, gravy granules, and flavored potato snacks). When yeast extracts are prepared from whole cells, residual cell wall material (referred to as yeast hulls or yeast glycan) is generated. This material has potential in the food industry (as a stabilizer, nonnutritive bulking agent, and emulsifier), the environment (as heavy-metal biosorbent material), and in medicine (as immunomodulators to protect against bacterial and fungal infections).

Other industrial uses of yeast biomass are as biotherapeutic agents (for example, *S. boulardii* has been reported to combat harmful intestinal bacteria) and biocatalysts (for example, baker's yeast is used by organic chemists in certain catalytic conversions such as reductions of aldehydes and ketones). Several yeasts have potential use as biocontrol agents to combat fungal crop disease. Another application is in bioremediation, in which yeasts are used to sequester heavy metals from industrial effluents. Immobilized yeast cells are used as biosensors for monitoring levels of certain aquatic pollutants. *See* BIODEGRADATION; INDUSTRIAL MICROBIOLOGY.

Industrial enzymes and chemicals. Although yeasts are not as rich a source of useful enzymes as certain bacteria or fungi, several enzymes prepared from yeast fermentations have applications in the food processing industries. Some of these enzymes are listed, along with several industrially useful yeast-derived chemicals, in **Table 2.** Genetic engineering of yeasts has the potential to increase the range of industrial enzymes and other commodities currently produced. For example, *S. cerevisiae* can be transformed by recombinant DNA technology to synthesize certain hydrolytic enzymes that enable this yeast to break down polymers, such as starch. Such technology would be very desirable in the conversion of starchy wastes into ethanol. *See* ENZYME; GENETIC ENGINEERING.

TABLE 2. Industrially useful enzymes and chemicals produced by yeasts

Commodity	Examples	Yeasts employed	Applications
Enzymes	Invertase (sucrose hydrolysis)	*Saccharomyces cerevisiae*	Confectionery, invert sugar
	Lactase (lactose hydrolysis)	*Kluyveromyces marxianus*	Lactose-reduced milk
	Lipase (triglyceride hydrolysis)	*Yarrowia lipolytica*	Infant foods, flavor modifier
	Protease (protein hydrolysis)	*Kluyveromyces lactis*	Cheese making
	Pectinase (pectin hydrolysis)	*Kluyveromyces* spp.	Wine/juice clarification
Organic acids	Citric acid	*Yarrowia lipolytica*	Foods, pharmaceuticals
	Malic acid	Food acidulant	
Fatty acids	Stearic acid	*Cryptococcus curvatus*	Cocoa butter equivalents
Amino acids	Lysine, tryptophan	*Candida, Saccharomyces*	Dietary supplements, therapy
Vitamins	Thiamin, riboflavin, biotin, pyridoxin, inositol, choline, nicotinc and folic acids	*Saccharomyces cerevisiae, Candida utilis*	Therapeutic uses
Sterols	Ergosterol	*Saccharomyces cerevisiae*	Precursor of vitamin D_2 (feeds)
Polysaccharides	Phosphomannans	*Pichia, Pachysolen, Hansenula*	Food emulsifiers
Polyols	Glycerol	*Saccharomyces cerevisiae*	Nitroglycerine explosives (historical)

Pharmaceutical products. Yeast preparations were used in early medicine to treat bacterial infections and heal wounds. More recently, yeasts have been administered as probiotics to combat harmful intestinal bacteria, and baker's yeast has been considered useful in the treatment of acne and premenstrual stress. The use of genetically manipulated yeasts in modern biotechnology for the production of novel pharmaceuticals is on the increase. Yeasts possess distinct advantages over bacteria for expression of foreign genes. The fact that yeasts are eukaryotic (as opposed to bacteria, which are prokaryotes) means that they have a cellular makeup similar to mammalian cells and can synthesize and process many pharmacologically useful proteins for the treatment and prevention of human disease. Several foreign genes have already been successfully expressed in yeast cells. Examples of therapeutic proteins produced from yeast include viral vaccines (prevention of hepatitis B), hormones (insulin for diabetics), blood proteins (human serum albumin for use in transfusions), growth factors (epidermal growth factor for wound healing), and interferons (beta-interferon for treatment of multiple sclerosis). Although *S. cerevisiae* has been considered a good host for foreign gene expression in this field, other yeasts such as *Pichia pastoris*, *Hansenula polymorpha*, and *Kluyveromyces lactis* have great potential in the production of natureauthentic, bioactive therapeutic proteins using recombinant DNA technology. *See* BIOCHEMICAL ENGINEERING; BIOTECHNOLOGY.

Graeme M. Walker

Bibliography. J. R. Dickinson and M. Schweizer (eds.), *The Metabolism and Molecular Physiology of Saccharomyces cerevisiae*, Taylor and Francis, London, 1998; W. M. Ingledew, Alcohol production by *Saccharomyces cerevisiae*: A yeast primer, in K. A. Jacques et al. (eds.), *The Alcohol Textbook*, 3d ed., Nottingham University Press, 1999; C. P. Kurtzman and J. W. Fell, *The Yeasts: A Taxonomic Study*, 4th ed., Elsevier, Amsterdam, 1998; N. P. Money, More g's than the Space Shuttle: Ballistospore discharge, *Mycologia*, 90(4):547–558, 1998; H. J. Phaff et al., *The Life of Yeasts*, rev. ed., 1978; G. Reed (ed.), *Prescott and Dunn's Industrial Microbiology*, 4th ed., 1982; F. A. Skinner et al. (eds.), *Biology and Activities of Yeasts*, 1981; J. F. T. Spencer and D. M. Spencer (eds.), *Yeasts in Natural and Artificial Habitats*, Springer-Verlag, Berlin, 1997; J. F. T. Spencer et al. (eds.), *Yeast Genetics*: *Fundamental and Applied Aspects*, 1983; J. F. T. Spencer and D. M. Spencer, *Yeast Technology*, 1989; A. Vaughan and A. Martini, Facts, myths and legends on the prime industrial microorganism, *J. Ind. Microbiol.*, 14:514–522, 1995; G. M. Walker, *Yeast Physiology and Biotechnology*, Wiley, New York, 1998; A. E. Wheals, A. H. Rose, and J. S. Harrison (eds.), *The Yeasts*, 6 vols., 1969–1995; K. Wolf (ed.), *Non-conventional Yeasts in Biotechnology: A Handbook*, Springer-Verlag, Berlin, 1996; F. K. Zimmermann and K. D. Entian (eds.), *Yeast Sugar Metabolism*, Technomic Publishing AG, Basel, Switzerland, 1997.

Yellow fever

An acute, febrile, mosquito-borne viral disease characterized in severe cases by jaundice, albuminuria, and hemorrhage. Inapparent infections also occur.

Infectious agent. The agent is a flavivirus, an arbovirus of group B. The virus multiplies in mosquitoes, which remain infectious for life. After the mosquito ingests a virus-containing blood meal, an interval of 12–18 days (called the extrinsic incubation period) is required for it to become infectious. *See* ANIMAL VIRUS; ARBOVIRAL ENCEPHALITIDES.

The virus enters the body through a mosquito bite and multiplies in lymph nodes, circulates in the blood, and localizes in the liver, spleen, kidney, bone marrow, and lymph glands. The severity of the disease and the major signs and symptoms which appear depend upon where the virus localizes and how much cell destruction occurs. The incubation period is 3–6 days. At the onset, the individual has fever, chills, headache, and backache, followed by nausea and vomiting. A short period of remission often follows. On about the fourth day, the period of intoxication begins with a slow pulse relative to a high fever and moderate jaundice. In severe cases, there are high levels of protein in the urine, and manifestations of bleeding appear; the vomit may be black with altered blood; and there is an abnormally low number of lymphocytes in the blood. When the disease progresses to the severe stage (black vomit and jaundice), the mortality rate is high. However, the infection may be mild and go unrecognized. Regardless of severity, individuals either die or recover completely.

Diagnosis. Diagnosis is made by isolation of the virus from the serum obtained from an individual as early as possible in the disease and inoculated intracerebrally in mice, or by the rise in serum antibody. The antibody response can be either of two types, depending upon the individual's previous experience with group B arboviruses. In a primary infection, the antibodies, particularly the complement-fixing and neutralizing antibodies, are quite specific. In a secondary infection, the individual's serum may react so broadly with other group B viruses that a specific diagnosis is sometimes impossible. *See* ANTIBODY; COMPLEMENT-FIXATION TEST; NEUTRALIZATION REACTION (IMMUNOLOGY).

Epidemiology. There are two major epidemiological cycles of yellow fever: classical or urban epidemic yellow fever, and sylvan or jungle yellow fever. Urban yellow fever involves person-to-person transmission by *Aedes aegypti* mosquitoes in the Western Hemisphere and West Africa. This mosquito breeds in the accumulations of water that accompany human settlement. Mosquitoes remain close to houses and become infected by biting a viremic individual. Urban yellow fever is perpetuated in areas where there is a constant influx of susceptible persons, some cases of yellow fever, and *A. aegypti.* With the use of intensive measures for mosquito abatement, the incidence of urban yellow fever has been markedly reduced in

South America, even though 200–400 cases are recognized annually, mainly in persons occupationally exposed in forested areas. The disease is probably underreported in Africa; epidemics involving forest mosquito vectors affect tens of thousands of humans at intervals of a few years, but only a few cases are officially reported.

Jungle yellow fever is primarily a disease of monkeys. In South America and Africa, it is transmitted from monkey to monkey by arboreal mosquitoes (*Haemagogus* and *Aedes* species) that inhabit the moist forest canopy. The infection in animals ranges from severe to inapparent. Persons, such as woodcutters or road builders, who come in contact with these mosquitoes in the forest can become infected. Jungle yellow fever may also occur when an infected monkey visits a human habitation and is bitten by *A. aegypti*, which then transmits the virus to a human.

Large numbers of inapparent infections occur. Yellow fever has never been reported in India or the East, even though the vector, *A. aegypti*, is widely distributed there.

New outbreaks continue to occur in other regions. In Bolivia, 145 cases of jungle yellow fever, with a mortality rate over 50%, were reported in 1975; and a large epidemic in Africa occurred in 1986 in Nigeria, with total of 3291 cases and 623 deaths. However, the actual incidence has been estimated to have been nearly three times the reported figure, and the number of deaths tenfold higher. *Aedes africanus* appears to have been the most likely vector in this epidemic. Yellow fever is a zoonosis that is difficult to control, and is capable of causing unpredictable epidemics in human populations. Yellow fever in the Americas continues to present epidemiological features typical of its jungle cycle: most cases are in males aged 15–45 years and engaged in agricultural or forestry activities.

Control and prevention. Vigorous mosquito abatement programs have virtually eliminated urban yellow fever. In the United States, the last reported outbreak of yellow fever occurred in 1905. However, with the speed of modern air travel, the threat of a yellow fever outbreak exists where *A. aegypti* is present. Most countries insist upon proper mosquito control on airplanes, and vaccinations of all persons at least 10 days before arrival in or from an endemic zone. However, the yellow fever vaccination requirement for travelers entering the United States was eliminated in 1972.

An excellent attenuated live-virus vaccine is available, in the 17D strain. Vaccine is prepared in fertile chicken eggs and dispensed as a dried powder. It is a live virus which is rehydrated just before use, and must be kept cold. *See* VACCINATION.

Joseph L. Melnick

Bibliography. P. L. Bres, A century of progress in combatting yellow fever, *Bull. WHO*, 4:775–786, 1986; A. S. Evans (ed.), *Viral Infections of Humans: Epidemiology and Control*, 4th ed., 1997; T. D. Monath, Flaviviruses, *Fields Virology*, 4th ed., 2001; T. P. Monath, Yellow fever: A medically neglected disease, *Rev. Infect. Dis.*, 9:165–175, 1987; M. Theiler and H. H. Smith, Use of yellow fever modified by an in vitro cultivation for human immunization, *J. Exp. Med.*, 65:787–800, 1937.

Yersinia

A genus of bacteria in the Enterobacteriaceae family. The bacteria appear as gram-negative rods and share many physiological properties with related *Escherichia coli*. Of the 11 species of *Yersinia, Y. pestis, Y. enterocolitica*, and *Y. pseudotuberculosis* are etiological agents of human disease. *Yersinia pestis* causes flea-borne bubonic plague (the black death), an extraordinarily acute process believed to have killed over 200 million people during human history. Enteropathogenic *Y. pseudotuberculosis* and *Y. enterocolitica* typically cause mild chronic enteric infections. The remaining species either promote primary infection of fish (*Y. ruckeri*) or exist as secondary invaders or inhabitants of natural environments (*Y. aldovae, Y. bercovieri, Y. frederiksenii, Y. intermedia, Y. kristensenii, Y. mollaretii*, and *Y. rohdei*).

The three yersiniae pathogenic to humans share about 70-kilobase plasmids (termed pCD or pYV) encoding cytotoxins termed Yops and anti-inflammatory activities that are delivered to host cells via a type III secretion mechanism. *Yersinia pestis* possesses an additional approximately 100-kb plasmid (pMT) that facilitates survival in the flea vector plus an approximately 10-kb plasmid (pPCP) encoding a plasminogen activator that promotes tissue invasiveness associated with acute disease. Antibodies directed against antigens encoded by the plasmids [for example, pCD-encoded LcrV (V antigen) and pMT-encoded capsular (F1 antigen)] promote immunity against plague. The siderophore yersiniabactin is a chromosomally encoded virulence determinant that facilitates assimilation of iron in vivo by *Y. pestis, Y. pseudotuberculosis*, and highly virulent serotypes of *Y. enterocolitica*.

Plague in humans is typically diagnosed on the basis of clinical symptoms such as fever, septicemia, and formation of buboes (inflamed lymph nodes), and is confirmed upon biopsy by antibody examination or specific agglutination. In the absence of specific diagnostic reagents, *Y. pestis* can still be determined by the nature of its source, morphology, gram reaction, growth habit, and lack of motility as well as certain enzymes shared by other yersiniae. Fortunately, the pneumonic form of human plague seldom occurs at the immediate outset of an epidemic; thus, individuals exhibiting the initial symptoms of this almost uniformly fatal form of the disease are typically identified by the alerted physician and are subjected to prompt treatment with an effective antibiotic (for example, chloramphenicol, but not penicillin).

In North America, the phenomenon of sylvatic plague reflects the occurrence of a widespread reservoir of infected rodents of many species west of the Mississippi River. Contact between humans and infected fleas in these rural areas typically results in a modest number of fatalities almost every year

because of delay in seeking treatment, incorrect diagnosis, or use of ineffective antibiotics. *See* MEDICAL BACTERIOLOGY; PLAGUE.

Robert R. Brubaker

Bibliography. R. R. Brubaker, Factors promoting acute and chronic diseases caused by yersiniae, *Clin. Microbiol. Rev.*, 4:309–324, 1991; T. Butler, *Plague and Other Yersinia Infections*, Plenum Press, 1983; G. R. Cornelis et al., The virulence plasmid of *Yersinia*, an antihost genome, *Microbiol. Mol. Biol. Rev.*, 62:1315–1352, 1998; R. D. Perry and J. D. Fetherston, *Clin. Microbiol. Rev.*, 10:35–66, 1997.

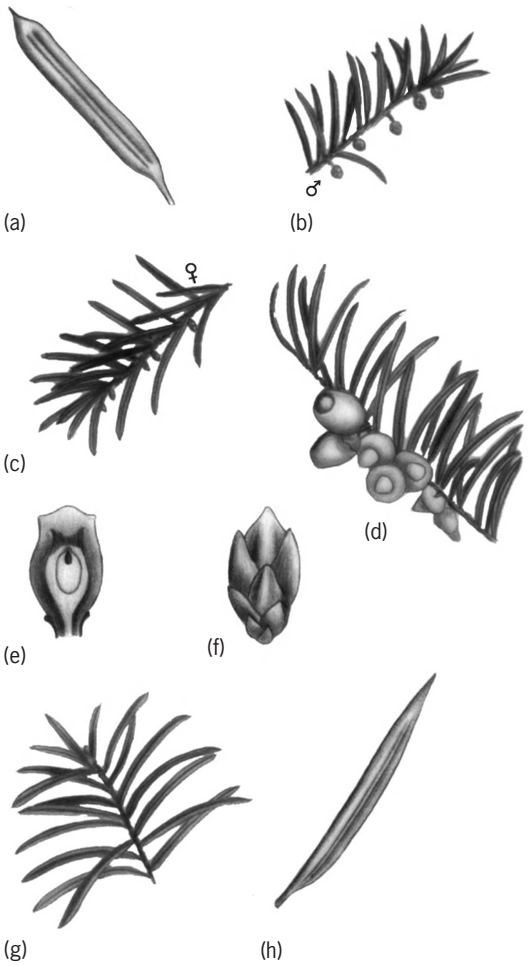
Yew

A genus of evergreen trees and shrubs, *Taxus*, with a fruit containing a single seed surrounded by a scarlet, fleshy, cuplike envelope (aril). The leaves are flat and acicular (needle-shaped), green below, with stalks extending downward on the stem. The only native American species of commercial importance is the rather uncommon Pacific yew (*T. brevifolia*), a medium-sized tree of the Pacific Coast and northern Rocky Mountain regions. Its wood is sometimes used for poles, paddles, bows, and small cabinetwork.

The English yew (*T. baccata*), native in Europe, North Africa, and northern Asia, and the Japanese yew (*T. cuspidata*) are much cultivated in the United States as evergreen ornamentals. Both are small trees when mature, but often in cultivation they are pruned to lower dimensions. In the English yew the leaves taper gradually to a point and are glossy, whereas in the Japanese yew the leaves are abruptly pointed and of duller appearance (see **illus.**). There are many cultivated forms of these two species.

The Canada yew (*T. canadensis*), also known as ground hemlock, is a low, straggling shrub of the forests of the northeastern quarter of the United States. It is distinguishable from a true hemlock by the absence of white lines on the underside of the leaves. *See* FOREST AND FORESTRY; TREE.

Arthur H. Graves; Kenneth H. Davis



Identifying characteristics of yews. (*a*) Japanese yew (*Taxus cuspidata*) leaf, (*b*) male branchlet with staminate flower buds, (*c*) female branchlet showing ovule-bearing flowers, (*d*) female branchlet with various stages of development of fleshy aril around seed, and (*e*) lengthwise section of berry showing seed in aril. (*f*) English yew (*T. baccata*) bud, showing obtuse, scarcely keeled scales, (*g*) branchlet, and (*h*) leaf.

Ylide

A variety of organic compounds which contain two adjacent atoms bearing formal positive and negative charges, and in which both atoms have full octets of electrons. Heteroatoms most commonly utilized as the positive atom are phosphorus, nitrogen, sulfur, selenium, and oxygen, and the negative atom usually involves carbon, nitrogen, oxygen, or sulfur. The ylide may be in an alicyclic or cyclic environment, and in the latter the ylide function may be endocyclic or exocyclic. Ylides most useful in organic synthesis are those containing phosphorus or sulfur and an adjacent carbanion.

Alicyclic and some cyclic ylides are most frequently prepared by deprotonation of the corresponding salt with a strong base. Thus, a phosphorus ylide (also called a phosphorane) is obtained from the phosphonium salt with *n*-butyllithium, the salt itself being prepared from the phosphine and an alkyl halide as shown in the reaction below, where

$$(Ph)_3P + \underset{\text{Alkyl halide}}{XCHR'R''} \rightarrow \underset{\text{Phosphonium salt}}{(Ph)_3\overset{\oplus}{P}CHR'R''\overset{\ominus}{X}} \xrightarrow{n\text{-BuLi}}$$

$$(Ph)_3P{=}CR'R'' \leftrightarrow (Ph)_3\overset{\oplus}{P}{-}\overset{\ominus}{C}R'R''$$

Phosphorus ylide

Ph = phenyl group, Bu = butyl group, and X = halogen. *See* REACTIVE INTERMEDIATES.

Ylides of this type are highly reactive, and when R′ and R″ are electron-withdrawing groups, the ylides are more stable. Their principal chemical application is reaction with an aldehyde or ketone, which may be alicyclic, aromatic, or heterocyclic, so that the resultant product contains a double bond at the site of the carbonyl group (the Wittig reaction). *See* ALDEHYDE.

Sulfur-containing ylides (π-sulfuranes) are of two types, sulfonium ylides and oxosulfonium ylides, which differ in having an oxygen atom attached to the sulfur in the latter. Both are usually prepared by deprotonation of the corresponding sulfonium

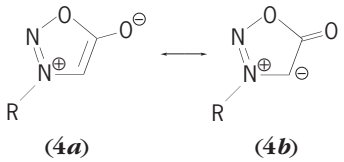
salts prepared by alkylation of the sulfide or sulfoxide with an active alkyl halide. Limitations on access to the sulfonium salts determine the availability of the ylides which may be stabilized by electron-withdrawing substituents on the carbon atom.

The addition of a carbene to a sulfide provides the most direct route to sulfur ylides and has given rise to stabilized ylides. Stabilized oxosulfonium ylides are specially suited to this approach. Structures (**1**)–(**3**) are examples of the sulfur-containing ylides.

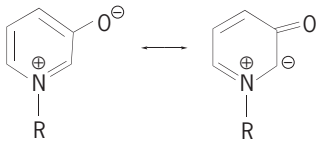
(**1**) (**2**) (**3**)

As with the phosphoranes, the main applications of sulfonium and oxosulfonium ylides are their reactions with aldehydes and ketones which lead to epoxides, and with α,β-unsaturated carbonyl compounds which provide a convenient method of introducing a cyclopropane ring.

The ylide function may also be incorporated into heterocyclic systems. For example, the meso-ionic sydnone molecule (***4a***) contains an azomethine imine ylide (***4b***), and the mesomeric betaine (***5a***) derived from 3-hydroxypyridine contains an azomethine ylide (***5b***). Both these ylides (often referred to



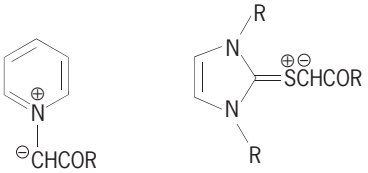
(***4a***) (***4b***)



(***5a***) (***5b***)

as masked ylides) readily undergo cycloaddition reactions formally analogous to the Diels-Alder reaction. *See* DIELS-ALDER REACTION; MESO-IONIC COMPOUND.

Other heterocyclic ylides are of the exocyclic type and usually are formed by proton abstraction from a heterocyclic salt. Structures (**6**) and (**7**) show two general types.



(**6**) (**7**)

Kevin T. Potts

Bibliography. F. A. Carey and R. J. Sundberg, *Advanced Organic Chemistry*, pt. A: *Structures and Mechanisms*, 4th ed., 2000; F. R. Hartley, *The Chemistry of Organophosphorus Compounds*, vol. 3: *Phosphonium Salts, Ylides, and Phosphoranes*, 1994; J. March, *Advanced Organic Chemistry: Reactions, Mechanisms, and Structures*, 5th ed., 2001; A. Padwa (ed.), *1,3-Dipolar Cycloaddition Chemistry*, 1984.

Yolk sac

An extraembryonic membrane which extends through the umbilicus in vertebrates. In some elasmobranchs, birds, and reptiles, it is laden with yolk which serves as the nutritive source of embryonic development. Developed in the splanchnopleure, it is composed of an endodermal layer and a mesodermal layer in which blood vessels are formed. *See* GERM LAYERS.

Chick embryo. The yolk sac develops as a membrane, from extraembryonic splanchnopleure, that grows over the surface of the yolk in embryos such as the chick. The yolk sac connects directly with the gut of the embryo at the midgut region. Yolk is broken down by enzymes present in the inner lining (endoderm) of the yolk sac membrane. Once digested, yolk passes through the yolk sac into blood vessels that are formed in the outer (mesoderm) layer of the yolk sac. Circulation carries the digested yolk to all parts of the developing embryo. As a chick approaches hatching, the yolk sac and remaining yolk become enclosed in the embryo proper. This yolk provides a newly hatched chick with an important source of nutrition.

Blood circulation. The yolk sac is the first source of blood and blood vessels in the chick embryo. Blood islands appear in the yolk sac and quickly differentiate into blood channels enclosing blood cells. The forming vessels in the blood islands extend and join, forming a vitelline plexus which joins larger vessels, the vitelline veins and arteries. These vessels join others and connect to the heart, which begins to pulsate, allowing circulation between the yolk sac and the embryo proper.

The circulation of blood from the yolk sac to the heart and embryo proper is relatively simple. Blood circulates in the small vessels of the yolk sac, picking up yolk components that have been liquefied by the digestive enzymes in the yolk sac endoderm. The blood eventually reaches one of the vitelline veins, which join the omphalomesenteric veins. These veins fuse into a single vein that enters the heart of the embryo proper. Blood from various sources mixes in the heart and leaves through the ventral aorta, then travels through the aortic arches into the dorsal aorta, and finally some of the blood returns to the yolk sac through the vitelline arteries.

Red blood cell production. The development of hemoglobin-rich red blood cells, which carry oxygen to different parts of the body, begins in the yolk sac. In the mouse embryo, for example, an embryonic form of hemoglobin begins to appear in the red blood cell precursors in the blood islands of the yolk sac by the eighth day of gestation. By the ninth day,

these cells are beginning to enter circulation. Later, the liver and then the spleen and bone marrow take over red blood cell production.

Mammals. In mammals, as in birds, the yolk sac generally develops from extraembryonic splanchnopleure, and extends beneath the developing embryo. Just as in the chick, a blood vessel network develops in the mammalian yolk sac lining. Though these blood vessels are empty, they play an important role in absorbing nourishing food and oxygen from the mother. Thus, although the yolk sac in higher mammals may be considered an evolutionary vestige from its yolky-egged ancestors, it still serves important functions in the young embryo. As the embryo ages, the yolk sac shrinks in size, and the allantois takes over the role of nutrition. *See* ALLANTOIS.

Role in immune system. The stem cells that give rise to the cells of the immune system in the chick embryo are also derived from the yolk sac wall. These cells, characterized by darkly staining basophilic cytoplasm, enter the two glands that are associated with the early development of the immune system, the thymus gland and the bursa of Fabricius. In mammalian embryos, such as the mouse, the stem cells of the immune system also originate in the yolk sac wall. These cells travel from the yolk sac to the fetal liver and then finally to the bone marrow.

Primordial germ cell origin. In some organisms, the yolk sac wall is also the site of origin of the primordial germ cells, those cells that eventually form the sperm and eggs. In mammals, the primordial germ cells appear to originate in the endoderm of the yolk sac. These cells contain large amounts of the enzyme alkaline phosphatase and therefore can be observed by using a special stain for this enzyme. The primordial germ cells migrate from the yolk sac wall, through adjoining dorsal mesentery, and into the genital ridges that form the gonads. *See* EMBRYOLOGY; FETAL MEMBRANE.

Steven B. Oppenheimer

Bibliography. H. Butler and B. H. Juvrlink, *An Atlas for Staging Mammalian and Chick Embryos*, 1987; B. M. Carlson, *Patten's Foundations of Embryology*, 6th ed., 1996; S. B. Oppenheimer and E. J. Carroll, *Introduction to Embryonic Development*, 4th ed., 2004; T. W. Sadler, *Langman's Medical Embryology*, 8th ed., 2000.

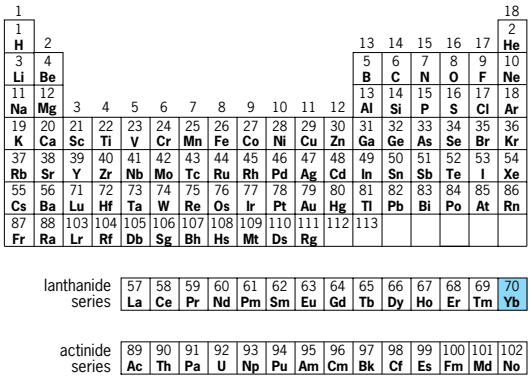
Young's modulus

A constant designated E, the ratio of stress to corresponding strain when the material behaves elastically. Young's modulus is represented by the slope $E = \Delta S/\Delta\epsilon$ of the initial straight segment of the stress-strain diagram. More correctly, E may be represented as the slope of the tangent or the slope of the secant connecting two points of the stress-strain curve. The modulus is then designated as tangent modulus or secant modulus at stated values of stress. The modulus of elasticity applying specifically to tension is called Young's modulus. Many materials have the same value in compression. *See* ELASTICITY; HOOKE'S LAW; STRESS AND STRAIN.

W. J. Krefeld; W. G. Bowman

Ytterbium

A chemical element, Yb, atomic number 70, and atomic weight 173.04. Ytterbium is a metal element of the rare-earth group. There are 7 naturally occurring stable isotopes. *See* PERIODIC TABLE.



The common oxide, Yb_2O_3, is colorless and dissolves readily in acids to form colorless solutions of trivalent salts which are paramagnetic. Ytterbium also forms a series of divalent compounds. The divalent salts are soluble in water but react very slowly with water to liberate hydrogen.

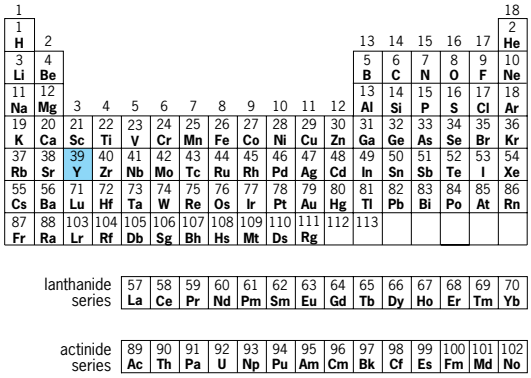
The metal is best prepared by distillation. It is a silvery soft metal which corrodes slowly in air and resembles the calcium-strontium-barium series more than the rare-earth series. For a discussion of the properties of the metal and its salts *See* RARE-EARTH ELEMENTS.

Frank H. Spedding

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; K. A. Gschneidner Jr., J.-C. Bünzli, and V. K. Pecharsky (eds.), *Handbook on the Physics and Chemistry of Rare Earths*, 2005.

Yttrium

A chemical element, Y, atomic number 39, and atomic weight 88.905. Yttrium resembles the rare-earth elements closely. The stable isotope ^{89}Y constitutes 100% of the natural element, which is always found associated with the rare earths and is



frequently classified as one. *See* PERIODIC TABLE.

Yttrium metal absorbs hydrogen, and in alloys up to a composition of YH_2 they resemble metals very closely. In fact, in certain composition ranges, the alloy is a better conductor of electricity than the pure metal.

Yttrium forms the matrix for the europium-activated yttrium phosphors which emit a brilliant, clear-red light when excited by electrons. The television industry uses these phosphors in manufacturing television screens.

Yttrium is used commercially in the metal industry for alloy purposes and as a "getter" to remove oxygen and nonmetallic impurities in other metals. For properties of the metal and its salts *see* RARE-EARTH ELEMENTS

Frank H. Spedding

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; M. F. Lappert, *Comprehensive Organometallic Chemistry II: Scandium, Yttrium, Lanthanides and Actinides, and Titanium, Zirconium, and Hafnium*, vol. 4, 1995.

Z transform — Zygophyllales

Z transform

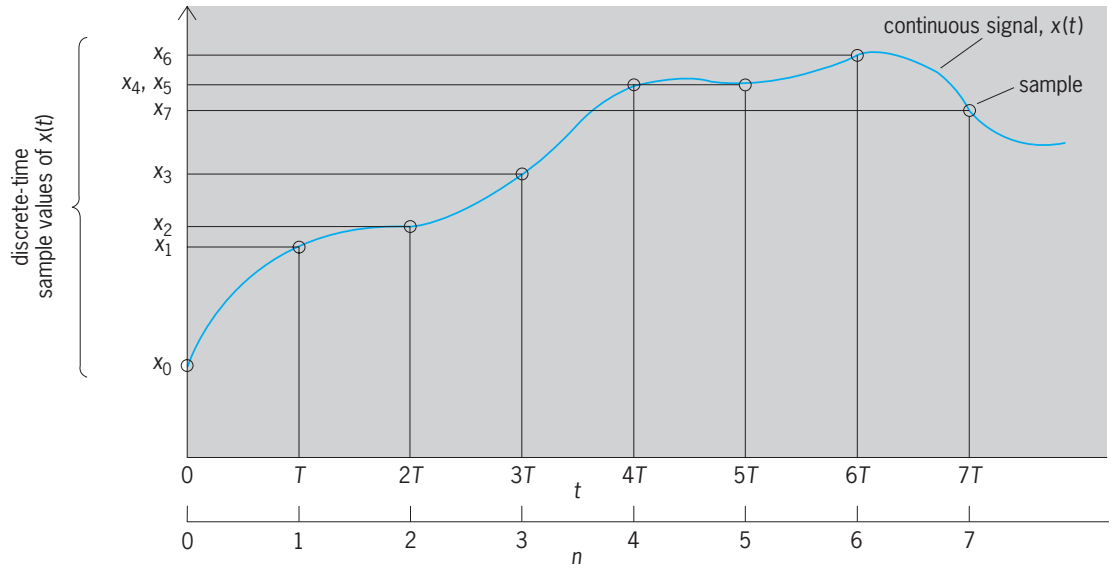
The preferred operational-calculus tool for analysis and design of discrete-time systems. (It should not be confused with the z transformation.) The role of the z transform with regard to discrete-time system is similar to that of the Laplace transform for continuous systems. In fact, the Laplace transform is a specialized case of the z transform. The z transform is by far the more insightful tool, and the Laplace transform is just the limiting case of the z transform in a practical as well as a conceptual way. *See* CONTROL SYSTEMS; DIGITAL FILTER; LAPLACE TRANSFORM; LINEAR SYSTEM ANALYSIS.

Definition. It is useful to consider a band-limited real continuous signal, $x(t)$, with no significant amount of energy above a frequency f_c. This signal is sampled at uniformly spaced intervals of time, T, $2T$, $3T, \ldots, nT, \ldots$ (see **illus.**), where the sampling interval, T, and the sampling frequency, f_s, are reciprocals of one another, and where $f_s > 2f_c$, a condition necessary for unambiguous interpretation of the sampled signal. If the common shorthand notation $x(nT) = x_n$ is used, the definition of the z transform of $x(t)$ is given by the equation below. The coefficient of

$$X(z) = x_0 + x_1 z^{-1} + x_2 z^{-2} + x_3 z^{-3} + \cdots$$

z^{-p} is therefore the value of the pth sample of the time signal. This gives a great deal of physical insight to the use of the z transform, a feature not shared by the Laplace transform. The Laplace transform can be found by evaluating the limit of the z transform as T approaches zero. *See* ANALOG-TO-DIGITAL CONVERTER; INFORMATION THEORY.

Poles, zeros, and the z plane. The z-transform expressions are polynomials which may be written in powers of either z or z^{-1}. The choice is a matter of personal preference. The quantity z^{-1} has the physical significance of representing one sample-period of delay. When the z-transform polynomials are



Extraction of discrete-time sample values from a continuous signal.

expressed in powers of z, the process of factoring and finding the roots is straightforward. The roots of the numerator and denominator are called zeros and poles, respectively, and are often mapped in the complex z plane. The vertical and horizontal axes of the complex z plane are the imaginary and real, respectively. Each pole location is traditionally marked by an x, and each zero location is marked by an o.

If the magnitudes of all poles are less than one (that is, they all lie within the unit circle), the corresponding time function will decrease to zero with increasing time. If there are one or more distinct poles whose magnitudes are one (that is, that sit on the unit circle), the corresponding time function will not go to zero with increasing time, but will be periodic and bounded. If any pole is greater than one (that is, if any one or more lie outside the unit circle), or if there are repeated (that is, identical) poles whose magnitudes are one (that is, those that sit on the unit circle), the corresponding time function will grow without bound as time increases. The locations of the zeros have no such effect.

The region of convergence in the z plane is that area in which poles may lie corresponding to a time function that decays to zero. For all positive time, the region of convergence is the area within the unit circle. *See* COMPLEX NUMBERS AND COMPLEX VARIABLES.

Stanley A. White

Bibliography. R. V. Churchill and J. W. Brown, *Complex Variables and Applications*, 6th ed., 1995; L. C. Ludeman, *Fundamentals of Digital Signal Processing*, 1986; A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 2d ed., 1999; A. V. Oppenheim and A. S. Willsky, *Signals and Systems*, 2d ed., 1996.

Zebra

One of three species of striped, horselike mammals belonging to the family Equidae and found wild in Africa. Zebras are classified in the order Perissodactyla (odd-toed ungulates), a group of mammals in which the middle toe is functional and the second and fourth digits are vestigial. All equids walk on the tips of their toes (unguligrade locomotion). Zebras stand 1.2 to 1.5 m (4–5 ft) high at the withers. Adults weigh 106–202 kg (235–450 lb). The dental formula is I 3/3, C 1/1, PM 3/3, M 3/3 × 2 for a total of 40 teeth.

Grevy's zebra, *Equus grevyi*; Samburu National Reserve, Kenya. (*Gerald and Buff Corsi; copyright © California Academy of Sciences*)

Large numbers of zebras once lived over most of the eastern part of Africa, from southern Egypt to the Cape of Good Hope. They were killed for their meat, which is said to have an excellent taste, and their hides, which are usd to make a tough leather. Some species are nearly extinct, but others are numerous.

Zebras differ from all other members of the horse family because of their startling color pattern. They have parallel black or dark brown stripes on a whitish background arranged in exact designs. These stripes run all over the body, meeting diagonally down the sides of the head. The lines may appear even on the zebra's long ears, short thick mane, and down its tail to the tuft of hair at the tip. Researchers have rejected the hypotheses that stripes serve as camouflage for zebras, that they visually confuse predators and pests, or that they assist in regulating body temperature through heat absorption. Instead, it is thought that stripes facilitate group cohesion and socialization.

Zebras are sociable grazing animals and may be found in association with gnu, deer, and ostriches. Some live on open grassy plains, whereas others live in rough mountains. They live in small bands or family groups, each of which is led by a stallion. Additional stallions live in small stallion groups. Zebras are difficult to tame and train to work.

Reproduction. Births may occur in all months, with gestation periods ranging from 360 days to 13 months. A single foal is the norm. Foals are up and about within an hour of their birth. They start grazing within a few weeks, but are generally not weaned for 8 to 13 months. Females come into heat 7–10 days after giving birth; thus mating and birth occur during the same season. Although females can breed annually, the interbirth interval ranges from 18 to 36 months because of the strain of rearing the foal. Captive zebras have survived for 40 years.

Species and distribution. The plains, or common, zebra (*Equus burchellii*) inhabits the grasslands, light woodlands, open scrub, and savannahs of East Africa south of the Sahara Desert. The largest population (over 250,000) occurs in the Serengeti of Tanzania. This species has a sleek body with broad vertical black and white stripes running onto the belly and becoming horizontal on the haunches. There are three subspecies. The mountain zebra (*E. zebra*) is confined to the mountainous grasslands of southwest Africa. It has narrower stripes than the plains zebra, broad black stripes on the rump, and a white belly. A dewlap (fold of skin) is present on the underside of the neck. The two subspecies, both endangered and close to extinction, are now found only in small numbers on protected reserves. Grevy's zebra (*E. grevyi*, see **illustration**), is the largest

living wild equine, reaching a height of 1.5 m (5 ft) at the withers. It inhabits subdesert steppe and arid bushy grasslands in Ethiopia, Somalia, and northern Kenya. It has narrow, vertical black and white stripes on its slender body, which curve upward on the haunches. The belly is white and unstriped, the mane is prominent and erect, and the legs are long. The head is long and narrow with prominent broad ears giving it a mulelike appearance. This species is also endangered.

The quagga (*E. quagga*) is an extinct southern African mammal that resembled a zebra. The last known individuals died in captivity in Berlin in 1875 and in Amsterdam in 1883. Some observers considered the quagga to be most closely related to the horse based on analyses of mainly cranial characters. Others thought it was a distinct species of zebra related to the three living species. Still others felt it was merely the southern end of a cline (graded series of characters) and a subspecies of the plains zebra. Both DNA and protein analyses of samples from quagga skin confirmed that it was, indeed, related to the plains zebra (*E. burchelli*). A breeding program is now under way in an attempt to "recreate" the quagga by repeated inbreeding of the most quaggalike plains zebras. Molecular studies may thus prove responsible, if only indirectly, for the return of one extinct subspecies (at least one that superficially resembles the quagga, since we can never be sure of the evolutionary pathways that created it) to its native habitat.

Zebra numbers continue to decline, especially on unprotected lands where humans bring settlement, crops, and livestock ranching. Natural predators include lions and hyaenas. *See* AFRICA; PERISSODACTYLA; EQUIDAE. Donald W. Linzey

Bibliography. D. W. Linzey, *Vertebrate Biology*, McGraw-Hill, 2000; D. Macdonald (ed.), *The Encyclopedia of Mammals*, Andromeda Oxford, 2001; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Zebu

A breed of humped, domestic cattle native to India and belonging to the family Bovidae in the order Artiodactyla. Also known as Brahman, they are probably descended from the wild auroch (*Bos taurus*), which is considered to be the ancestor of domestic cattle.

Zebus (*Bos indicus*) have been domesticated in Asia for approximately 6500 years. They were imported by breeders into the United States in 1849 and have been crossed with local breeds of beef cattle to produce strains better adapted to the hot humid Gulf states. The breed has also contributed to beef production though cross breeding with European cattle, such as Hereford and Angus. These cattle exhibit hybrid vigor, that is, they generally exhibit growth and reproductive rates greater than either of the parental types. Several new breeds of cattle have been developed in the United States based on Brahman-European crosses, some important ones being the Beefmaster (Brahman combined with Shorthorn cattle and Hereford cattle); Brangus (Brahman combined with Angus cattle); charbray (Brahman combined with Charolais cattle); and Santa Gertrudis (Brahman combined with Shorthorn). The Santa Gertrudis was the first distinct breed of cattle produced in the United States. It was developed in the 1920s and 1930s at the King Ranch in Kingsville, Texas, by a cross of Shorthorn with Brahman, and was first recognized as a distinct breed in 1940. Large numbers of them live in the humid areas of the South and Southwest. The calves grow rapidly and mature into large cattle. Santa Gertrudis have been exported to Africa and several Latin American countries.

Brahman (*Bos indicus*).

Brahman cattle have a very distinctive appearance with a large, fleshy, fatty hump (or sometimes double hump) over the shoulders, a rounded forehead, loose skin under the throat (dewlap), large drooping ears, and white legs (see **illustration**). The hump is formed from two overdeveloped muscles plus fatty tissue and probably represents an energy reserve for emergencies. These cattle are generally light to medium gray in color, but domestic animals may also be red, brown, or black. Brahman cattle have short hair and well-developed sweat glands that enable them to withstand heat and humidity, ticks, and insects.

Breeding may occur throughout the year, and a single offspring is born after a gestation of 277–290 days. Females attain sexual maturity at about 18 months and remain fertile for about 12 years. The life span may be 20+ years.

Zebus are protected as sacred cows by the Hindus in India, who allow them to roam freely through the streets and villages. In India, they are used as draft animals as well as for milk. Zebus spread from India to parts of China and as far west as East Africa. Some strains were developed for saddle riding. Others were bred for draft purposes or as pack animals. As a result of such selective breeding, zebus vary in weight and in size. Some may be as big as the largest ox. *See* ARTIODACTYLA; BEEF CATTLE PRODUCTION; BOVIDAE; CATTLE. Donald W. Linzey

Bibliography. *Grzimek's Encyclopedia of Mammals*, vol. 5, McGraw-Hill, 1990; R. M. Nowak, *Walker's Mammals of the World*, 6th ed., Johns Hopkins University Press, 1999.

Zeeman effect

A splitting of spectral lines when the light source being studied is placed in a magnetic field. Discovered by P. Zeeman in 1896, the Zeeman effect furnishes information of prime importance in the analysis of spectra. Each kind of spectral term has its characteristic mode of splitting, and the types of terms are most definitely identified by this property. Furthermore, the effect allows an evaluation of the ratio of charge to mass of the electron and an evaluation of its precise magnetic moment.

Normal Zeeman effect. This is a splitting into two or three lines, depending on the direction of observation, as shown in **Fig. 1**. The light of these

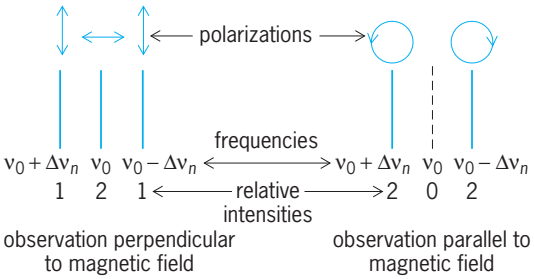


Fig. 1. Triplet observed in normal Zeeman effect.

components is polarized in ways indicated in the figure. The normal effect is observed for all lines belonging to singlet systems, those for which the spin quantum number S = 0. The change of frequency $\Delta\nu_n$ of the shifted components can be evaluated on classical electromagnetic principles as follows. Assume that the electron of charge e revolves in a circular orbit of radius r and circular frequency ω radians per second (**Fig. 2**). If a magnetic field H is applied perpendicular to the plane of the orbit, the electron will be speeded up or slowed down because of the changing flux through its orbit, just as in the electron accelerator known as the betatron.

Denoting the centripetal force holding the electron on its orbit before application of the field by f_0 (Fig. 2*b*), and the additional force due to the motion of the electron across the field by f_H (Fig. 2*a* and *c*), one has Eqs. (1). In Fig. 2*a*, where the two forces are

$$f_0 = m\omega^2 r \qquad f_H = He(\omega \pm \Delta\omega)r \tag{1}$$

in the same direction, one may equate their sum to the centripetal force on an electron of frequency $\omega + \Delta\omega$, obtaining Eq. (2). Solution of this equation for

$$f_0 + f_H = m\omega^2 r + He(\omega + \Delta\omega)r$$
$$= m(\omega + \Delta\omega)^2 r \tag{2}$$

$\Delta\omega$, under the assumption that it is small compared to ω itself, yields Eqs. (3), the latter expression following from the fact that $\nu = \omega/2\pi$. This relation, although derived for a special case, is generally valid for any system of particles having a particular value of e/m and moving under the action of a central force. *See* LARMOR PRECESSION.

$$\Delta\omega = \frac{eH}{2m} \qquad \Delta\nu_n = \frac{eH}{4\pi m} \tag{3}$$

On substitution of the radio of charge to mass of an electron, one obtains Eq. (4). Conversely, from

$$\Delta\nu_n = 1.3996 \times 10^6 H \text{ s}^{-1} \tag{4}$$

the observed spectroscopic splitting $\Delta\nu_n$ and measurement of the field strength, the value of e/m for the electron has been evaluated as 1.7572 ± 0.0007 emu/g. This is in good agreement with the figure determined by other methods.

Anomalous Zeeman effect. This effect is a more complicated type of line splitting, so named because it did not agree with the predictions of classical theory. It occurs for any spectral line arising from a combination of terms of multiplicity greater than one. As examples, **Fig. 3** gives diagrams of the theoretical patterns for the yellow lines of sodium, belonging to a doublet system, while **Fig. 4** shows some actual patterns observed for doublets and quartets in rhodium.

Since multiplicity in spectral lines is caused by the presence of a resultant spin vector S of the electrons, the anomalous effect must be attributed to a nonclassical magnetic behavior of the electron spin. While classical theory associates with the vector L of the orbital angular momentum a magnetic moment as in Eq. (5), it is necessary, in explaining the

$$\mu_L = \frac{eh}{4\pi mc} L \tag{5}$$

anomalous Zeeman effect, that the magnetic moment corresponding to S be as in Eq. (6). Thus the

$$\mu_S = \frac{eh}{2\pi mc} S \tag{6}$$

spin generates twice as much magnetic moment, relative to its angular momentum, as does the orbital motion. In an atom for which both L and S are finite, the effective magnetic moment may be written as Eq. (7), where J measures the total angular momentum (in LS coupling, the resultant of L and S), and μ_0 is the Bohr magneton, $eh/4\pi mc$. *See* PARAMAGNETISM.

$$\mu_J = g\frac{eh}{4\pi mc} J = g\mu_0 J \tag{7}$$

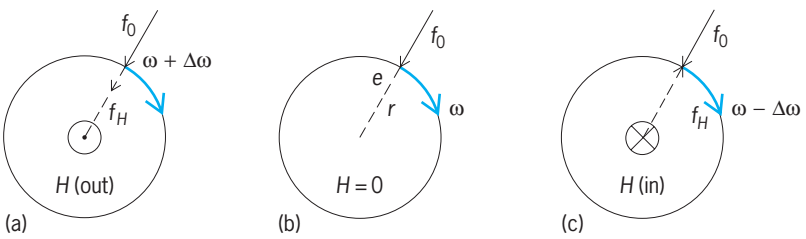


Fig. 2. Effect of a magnetic field *H* applied perpendicular to a circular electron orbit. Diagrams *a*–*c* are explained in the text.

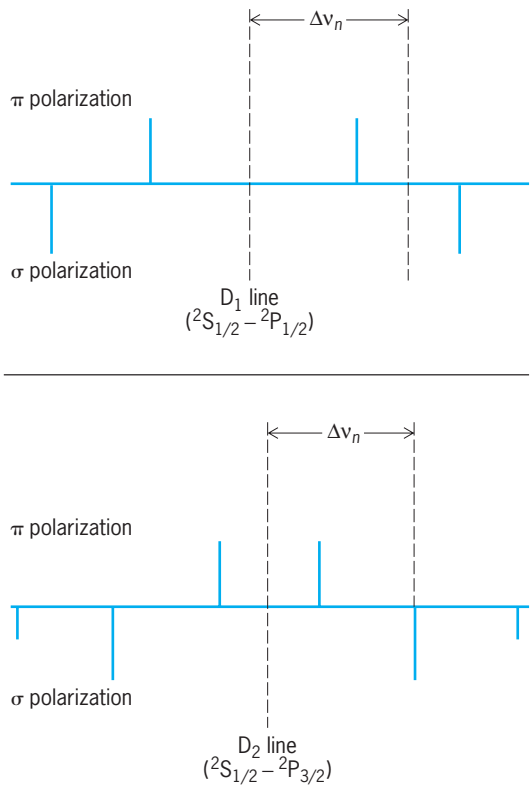


Fig. 3. Anomalous Zeeman effect of the sodium lines. $\Delta\nu_n$ denotes the normal Zeeman splitting, while π and σ refer to polarizations like those of the central and outer components in the normal effect illustrated in Fig. 1. The heights of the lines indicate their relative intensities.

Theory gives values of g, the Landé g factor, which are characteristic of the type of spectral term. In LS coupling, the value is given by Eq. (8).

$$g = 1 + \frac{J(J+1) + S(S+1) - L(L+1)}{2J(J+1)} \qquad (8)$$

For a classical electron orbit $g = 1$, which yields the normal Zeeman effect. When the spin S is present, however, the changes in energy produced by the magnetic field, which are proportional to μ_J, are just g times as great, and this fact is responsible for the anomalous Zeeman effect. It also should be mentioned that both theory and experiment now show that the g factor for the electron is not exactly 2, but 2.00229.

The component of μ_J in the field direction is $g\mu_0 M$, where M is the quantized component of J in this direction. The energy terms become $T = T_0 + g\mu_0 M$, where T_0 is the term value with no field. This magnetic quantum number M has only the $2J + 1$ values, $J, J - 1, J - 2, \ldots, -J$, and the allowed transitions between energy terms must obey the selection rule $\Delta M = 0, \pm 1$. To explain the statement that the normal Zeeman effect is observed for all lines in singlet systems, recall that S is then zero, $J = L$, and hence $g = 1$.

Quadratic Zeeman effect. The quadratic effect, which depends on the square of the field strength, is of two kinds. The first results from the second-order terms that were neglected in the preceding derivation, and the second, from the diamagnetic reaction of the electron when revolving in large orbits.

Inverse Zeeman effect. This is the Zeeman effect of absorption lines. It is closely related to the Faraday effect, the rotation of plane-polarized light by matter situated in a magnetic field. *See* ATOMIC STRUCTURE AND SPECTRA; FARADAY EFFECT.

Zeeman effect in molecules. This effect is, in general, so small as to be unobservable, even for molecules which have a permanent magnetic moment. Each level with a total angular momentum J splits into $2J + 1$ components, as in the case of atoms.

The component of the magnetic moment along the direction of the external field is small, however, because the rotation of the molecule, which carries the magnetic moment along with it, causes the principal part of the magnetic moment to average out to zero. The consequence is that the magnetic levels have an extremely narrow spacing except for cases where the molecule has either very little rotation or none at all. An exception occurs for some light molecules where the magnetic moment is coupled so lightly to the frame of the molecule that it can orient itself freely in the magnetic field just as for atoms (**Fig. 5**).

Zeeman effect in crystals. A clear Zeeman effect also can be observed in many crystals with sharp spectrum lines in absorption or fluorescence. Such crystals are found particularly among the salts of the rare earths. In these cases the internal electric field

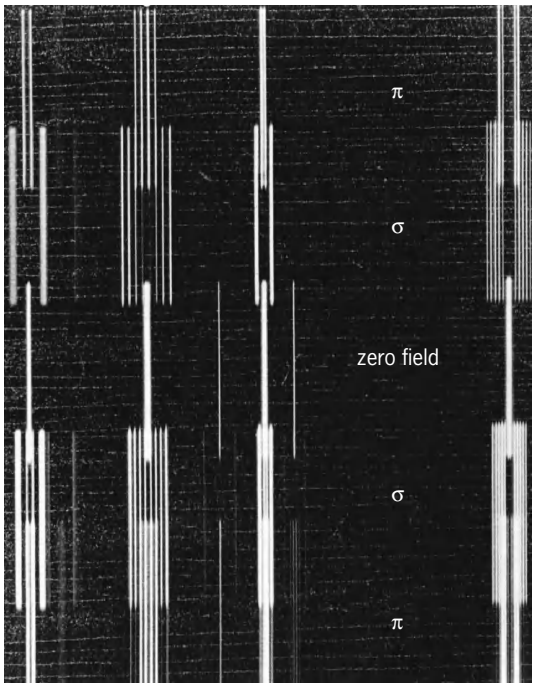


Fig. 4. Zeeman effect of the rhodium spectrum in the wavelength range 347.9–346.2 nanometers. Field strengths are 70,000 oersteds (lower exposure) and 90,500 (upper exposure). (*From G. R. Harrison and F. Bitter, Massachusetts Institute of Technology*)

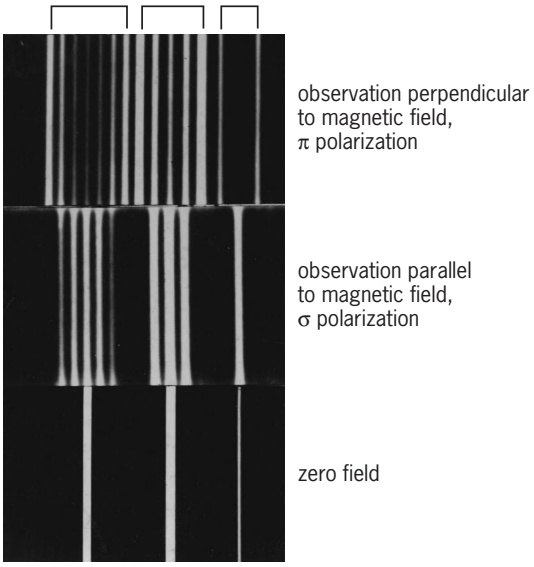


Fig. 5. Zeeman effect of three adjacent rotational lines of the hydrogen molecule. (*Johns Hopkins University*)

in the crystal splits and shifts the level of the free ion. When the number of electrons is even and the crystal symmetry low, this electric splitting is complete. No degeneracy remains, and there can be no further splitting by a magnetic field. If the number of electrons is odd, or if for an even number there is high crystal symmetry, the levels occur in degenerate pairs which are split by a magnetic field. Each line is then split into four components (**Fig. 6**). For cubic crystal symmetry and when angular momentum is due only to electron spin, splitting into more than four components may occur.

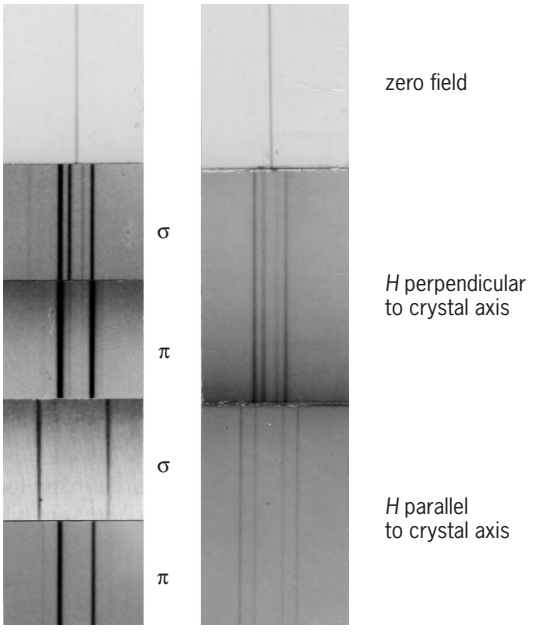


Fig. 6. Zeeman effect of an absorption line of neodymium chloride, with polarization (left) and without polarization (right). The splitting is quite different, depending on whether the trigonal crystal axis is parallel or perpendicular to the magnetic field. *H* = 35,000 oersteds. (*Johns Hopkins University*)

Nuclear Zeeman effect. The magnetic moment of the nucleus causes a Zeeman splitting in atomic spectra which is of an order of magnitude a thousand times smaller than the ordinary Zeeman effect. This Zeeman effect of the hyperfine structure usually is modified by a nuclear Paschen-Back effect, first studied by E. Back and S. A. Goudsmit for the spectral lines of bismuth. *See* PASCHEN-BACK EFFECT.

A strong magnetic field actually may modify the intensity and selection rules so that usually absent lines may appear. For example, the *J* selection rule is no longer valid in a magnetic field. *See* SELECTION RULES (PHYSICS). F. A. Jenkins; G. H. Dieke; W. W. Watson

Bibliography. R. D. Cowan, *The Theory of Atomic Structure and Spectra*, 1981; G. Herzberg, *Atomic Spectra and Atomic Structure*, 2d ed., 1944; F. A. Jenkins and H. E. White, *Fundamentals of Optics*, 4th ed., 1976; S. Svanberg, *Atomic and Molecular Spectroscopy*, 3d ed., 2001; L. Szasz, *The Electronic Structure of Atoms*, 1991; M. R. Wehr, J. A. Richards, and T. W. Adair, *Physics of the Atom*, 4th ed., 1984.

Zeiformes

The order of teleost fishes collectively called dories. The order has representatives in all seas of the world from the tropics to Antarctica; however, they are most common in tropical and temperate zones, in waters of moderate depth to at least 1800 m. Most species are mesopelagic or bathypelagic. Structurally intermediary to the beryciform and the perciform fishes, zeiforms are characterized by a body varying from moderately elongate to very deep and compressed; jaws usually greatly protractile; colors predominantly silvery or silver gray; absence of an orbitosphenoid bone; a simple posttemporal bone rigidly united to the skull; a pelvic fin each with one spine or none and from two to ten soft rays; a first dorsal fin with five to ten spines, usually strong; and an anal fin with one to four spines. Most species are of small size and of minor economic importance.

The order consists of five families. The Oreosomatidae have a very deep and compressed body, a superior and protractile mouth, and small cycloid or ctenoid scales; the young bear conical scutes on parts of body. The Parazenidae have a moderately deep body, highly protractile premaxillaries, weakly ctenoid scales that are not vertically elongate, two lateral lines that are fused behind the soft dorsal fin, and thoracic pelvic fins. This family consists of a single species, *Parazen pacificus*, known only from Japan and halfway around the world in waters off Cuba. Species of Zeniontidae (formerly Macrurocyttidae) have an elongate body, an extremely protractile upper jaw, large eyes, and a pelvic fin that may consist of a single large serrated spine preceding two inconspicuous soft rays and strong and serrated dorsal spines, with one quite larger than the others. Grammicolepididae have a very deep body, a small mouth with nearly vertical jaws, and vertically elongate scales. The Zeidae have a large, almost vertical mouth, small scales (if present) that

are not vertically elongate, and spines or bucklers (enlarged body scales) at the base of dorsal and anal fins. Herbert Boschung

Bibliography. P. C. Heemstra, A revision of the zeid fishes (Zeiformes: Zeidae) of South Africa, *Ichthyol. Bull. J. L. B. Smith Inst. Ichthyol*, 40:17, 1980; P. C. Heemstra, Zeniontidae, p. 441, in M. M. Smith and P. C. Heemstra (eds.), *Smith's Sea Fishes*, Springer-Verlag, Berlin, 1986; P. C. Heemstra, Zeidae. Dories, pp. 1207–1209, in K. E. Carpenter (ed.), *FAO Species Identification Guide for Fishery Purposes. The Living Marine Resources of the Western Central Atlantic*, vol. 2: *Bony Fishes, part 1 (Acipenseridae to Grammatidae)*, U.N. Food and Agricultural Organization (FAO), 2002.

Zener diode

A two-terminal semiconductor junction device with a very sharp voltage breakdown as reverse bias is applied. The device is used to provide a voltage reference. It is named after C. Zener, who first proposed electronic tunneling as a mechanism of electrical breakdown in insulators. *See* SEMICONDUCTOR; TUNNELING IN SOLIDS.

In the Zener diode, the normal rectifying characteristic is of no interest. Both the normal diode and the Zener diode voltage-current characteristics are shown in **Fig. 1**. When the Zener diode is reverse biased and is fed with a current from a higher potential source, a well-defined voltage is developed which mainly depends on the stability of the flowing current and the stability of the device temperature.

A classic circuit to define a very stable current uses an operational amplifier and three stable resistors (**Fig. 2**). The voltage across the Zener itself defines a higher level from which the current is drawn. Thus, a stable noise-free Zener defines its own stable noise-free current. *See* OPERATIONAL AMPLIFIER.

The effect of temperature on the breakdown voltage can be nulled by having a second forward-biased junction, which has a small negative temperature coefficient, in series with the Zener junction. Such a device is called compensated Zener and has a breakdown voltage of 6.2 V rather than the normal 5.6 V (for the smallest possible temperature coefficient). Alternatively, a Zener junction can be part of an integrated circuit which adds a whole temperature controller to keep the silicon substrate at a constant temperature. For the very best performance, only four components are integrated into the silicon: the Zener, a heater resistor, a temperature-sensing transistor, and a current-sensing transistor. A separate selected dual operational amplifier then completes the current-and-temperature-control circuit. Such a circuit sets the chip temperature at, say 122°F (50°C), and the junction condition is then largely independent of ambient temperature. *See* INTEGRATED CIRCUITS.

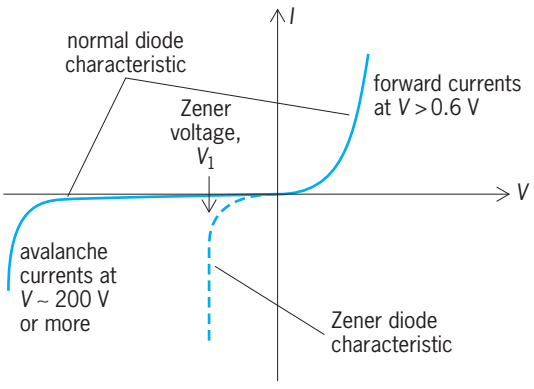


Fig. 1. Diode voltage-current (*V-I*) characteristics.

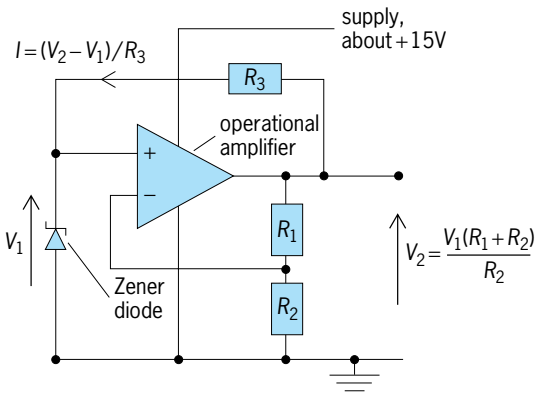


Fig. 2. Zener diode circuit to define a very stable current, *I*, using an operational amplifier and three stable resistors, R_1, R_2, and R_3. The voltage V_1 across the Zener itself defines a higher level V_2 from which the current *I* is drawn.

The *p* and *n* regions of the diode are often deposited under the silicon surface by ion bombardment. Then any surface contaminants will not affect the actual junction region. They would make the Zener voltage more noisy or less stable with time. *See* ION IMPLANTATION.

With the precautions mentioned to control the current and the temperature, the Zener voltage is still found to change by about 0.0002% per year. This is thought to be a function of the junction doping profiles slowly changing by diffusion. This rate of change of Zener voltage approximately doubles for each 18°F (10°C) that the device temperature is raised. Thus, the chip temperature must not be raised excessively to achieve temperature control.

The compact, robust Zener diode, with the circuits described to set its current and temperature, makes a fine portable voltage standard. This is used to disseminate the voltage level from national or accredited calibration laboratories to industry and to research laboratories. *See* JUNCTION DIODE; SEMICONDUCTOR DIODE; VOLTAGE MEASUREMENT; VOLTAGE REGULATOR. Peter J. Spreadbury

Bibliography. J. J. Brophy, *Basic Electronics for Scientists*, 5th ed., McGraw-Hill, 1990; P. Horowitz and W. Hill, *The Art of Electronics*, 2d ed., Cambridge University Press, 1989; J. Millman and A. Grabel, *Microelectronics*, 2d ed., McGraw-Hill, 1987; A. S. Sedra and K. C. Smith, *Microelectronic Circuits*, 4th ed., Oxford University Press, 1997; P. J. Spreadbury, The Ultra-Zener: A portable replacement for the Weston cell?, *IEEE Trans. Instrum. Meas.*, 40:343–346, 1991.

Zenith

The point in the sky directly above an observer. If the observer is at a pole, then the observer's zenith is a celestial pole; but for observers at midlatitudes the zenith is a point in the sky that corresponds to a changing right ascension but a constant declination as the sky rotates overhead. The point directly below the observer is the observer's nadir, and is 180° in longitude and in latitude from the observer's zenith.

The zenith distance of an object in the sky is the angle across the sky from the zenith to the object. The Sun is never overhead for observers outside the equatorial zone marked by the Tropic of Cancer and the Tropic of Capricorn, $\pm 23^1/_2^\circ$, so the zenith distance is never less than $21^1/_2^\circ$ for an observer at a latitude of 45°. The zenith distance is 90° minus the altitude of an object above the horizon. *See* ASTRONOMICAL COORDINATE SYSTEMS; CELESTIAL SPHERE; TROPIC OF CANCER; TROPIC OF CAPRICORN. Jay M. Pasachoff

Bibliography. J. Mitton, *The Penguin Dictionary of Astronomy*, 3d ed., Penguin, New York, 2000; J. M. Pasachoff and A. Filippenko, *The Cosmos: Astronomy in the New Millenium*, 3d ed., Brooks/Cole Publishing, Davis, California, 2007; P. K. Seidelmann (ed.), *Explanatory Supplement to the Astronomomical Almanac*, University Science Books, Mill Valley, CA, 1992.

Zeolite

Any mineral belonging to the zeolite family of minerals and synthetic compounds characterized by an aluminosilicate tetrahedral framework, ion-exchangeable large cations, and loosely held water molecules permitting reversible dehydration. The general formula can be expressed as $X_y^{1+,2+}Al_x^{3+}Si_{1-x}^{4+}O_2 \cdot nH_2O$. Since the oxygen atoms in the framework are each shared by two tetrahedrons, the (Si,Al):O ratio is exactly 1:2. The amount of large cations (X) present is conditioned by the aluminum/silicon (Al:Si) ratio and the formal charge of these large cations. Typical large cations are the alkalies and alkaline earths such as sodium (Na^+), potassium (K^+), calcium (Ca^{2+}), strontium (Sr^{2+}), and barium (Ba^{2+}). The large cations, coordinated by framework oxygens and water molecules, reside in large cavities in the crystal structure; these cavities and channels may even permit the selective passage of organic molecules. Thus, zeolites are extensively studied from theoretical and technical standpoints because of their potential and actual use as molecular sieves, catalysts, and water softeners. *See* MOLECULAR SIEVE; SILICATE MINERALS; WATER SOFTENING.

Occurrence. Zeolites are low-temperature and low-pressure minerals and commonly occur as late minerals in amygdaloidal basalts, as devitrification products, as authigenic minerals in sandstones and other sediments, and as alteration products of feldspars and nepheline. Phillipsite and laumontite occur extensively in sediments on the ocean floor. Stilbite, heulandite, analcime, chabazite, and scolecite are common as large crystals in vesicles and cavities in the basalts of the Minas Basin Region, Nova Scotia; West Paterson, New Jersey; the Columbia River Plateau; Berufjord, Iceland; Poona, India; and many other localities.

Zeolites are usually white, but often may be colored pink, brown, red, yellow, or green by inclusions; the hardness is moderate (3–5) and the specific gravity low (2.0–2.5) because of their rather open framework structures. Their habits are highly variable and depend on the atomic arrangement; some zeolites are fibrous and woolly while others are platy and micaceous and a few are equant in development and lacking in cleavage.

Crystal structure. As a result of detailed crystal structure analysis, the atomic arrangements of most zeolites are known. Important features are loops of 4-, 5-, 6-, 8-, and 12-membered tetrahedral rings which further link to form channels and cages. On the basis of the kinds of loops and channels present, the zeolite minerals can be classified according to groups (see **table**). The volume-per-framework oxygen atom is a measure of the packing efficiency of the structure, and zeolites have very high values, usually greater than 2.5 cubic nanometers (nm^3) per oxygen atom. This can be compared with a typical dense-packed silicate structure, such as olivine, which has a value of about 1.8 nm^3.

Knowledge of the crystal structure also explains the sieving properties of zeolites. The minimum width of a channel is an approximate measure of the maximum diameter of a molecule which can pass through. Thus, chabazite, with a channel of 0.39-nm minimum width, allows the passage of straight-chain hydrocarbons but not those with branched chains. Faujasite, with a 12-membered ring and channel diameter of 0.9 nm, even permits the passage of benzene rings. This substance, once a mineralogical curiosity, is now of key importance in the catalytic cracking of petroleum. Large quantities of faujasite and other zeolites are made synthetically, and their desirable qualities can be enhanced by tailoring their compositions.

Zeolites are usually synthesized hydrothermally, starting with gels of appropriate composition in an alkaline environment. The temperature of synthesis may range from 450 to 100°C (840 to 212°F) with the more open framework structures occurring at the lower temperatures. The appearance of zeolites in a metamorphic rock indicates a rock of lowest grade; such assemblages are referred to as the zeolite facies. *See* CRYSTAL STRUCTURE; FACIES (GEOLOGY). Paul B. Moore

Applications. The porous, yet crystalline nature of zeolites has been exploited commercially in three main areas—as sorbents, as cation-exchange materials, and as catalysts.

Sorbents. Most aluminosilicate zeolites have an extremely high (yet reversible) affinity for water and are widely used as desiccants. Zeolite A, in its potassium-exchanged form (K-A), is the material of choice for intensive drying of unsaturated hydrocarbon gas streams and polar liquids, where water contents as

Zeolite family of minerals

Group	Name	Formula	Crystal system	Volume/framework oxygen
Analcime	Analcime*	$Na(AlSi_2O_6) \cdot H_2O$	Cubic, pseudocubic	26.8
	Wairakite	$Ca(AlSi_2O_6) \cdot H_2O$	Monoclinic	
	Pollucite	$Cs(AlSi_2O_6) \cdot xH_2O$	Tetragonal	
Sodalite	Sodalite*	$Na_4(Al_3Si_3O_{12})Cl$	Cubic	30.0
	Linde A	$Na_{12}(Al_{12}Si_{12}O_{48}) \cdot 27H_2O$	Cubic	35.1
	ZK-5	$Na_{24}(Al_{24}Si_{72}O_{192}) \cdot 90H_2O$	Cubic	34.0
	Faujasite	$(Na_2,Ca)_{30}((Al,Si)_{192}O_{384}) \cdot 260H_2O$	Cubic	39.2
Chabazite	Chabazite*	$Ca_2(Al_4Si_8O_{24}) \cdot 13H_2O$	Rhombohedral, pseudorhombohedral	34.2
	Gmelinite	$Na_2(Al_2Si_4O_{12}) \cdot 6H_2O$	Hexagonal	33.8
	Erionite, offretite	$Ca_{4.5}(Al_9Si_{27}O_{72}) \cdot 27H_2O$	Hexagonal	31.9
	Levynite	$Ca(Al_2Si_4O_{12}) \cdot 6H_2O$	Rhombohedral	32.5
Natrolite	Natrolite*	$Na_2(Al_2Si_3O_{10}) \cdot 2H_2O$	Orthorhombic	28.3
	Scolecite	$Ca(Al_2Si_3O_{10}) \cdot 3H_2O$	Monoclinic	28.6
	Mesolite	$Na_2Ca_2(Al_2Si_3O_{10})_3 \cdot 8H_2O$	Monoclinic	
	Edingtonite	$Ba(Al_2Si_3O_{10}) \cdot 3H_2O$	Orthorhombic	30.1
	Thomsonite	$NaCa_2(Al_5Si_5O_{20}) \cdot 6H_2O$	Orthorhombic	28.3
	Gonnardite	$(Ca,Na)_{6-8}((Si,Al)_{20}O_{40}) \cdot 12H_2O$	Orthorhombic	
Phillipsite	Phillipsite	$(K,Na)_5(Al_5Si_{11}O_{32}) \cdot 10H_2O$	Orthorhombic	31.3
	Harmotome	$Ba_2(Al_4Si_{12}O_{32}) \cdot 12H_2O$	Monoclinic	30.9
	Gismondine	$Ca(Al_2Si_2O_8) \cdot 4H_2O$	Monoclinic	32.5
	Garronite	$NaCa_{2.5}(Al_6Si_{10}O_{32}) \cdot 13H_2O$	Tetragonal?	
Mordenite	Mordenite	$Na(AlSi_5O_{12}) \cdot 3H_2O$	Orthorhombic	29.2
	Dachiardite	$(Na_2,Ca)_2(Al_4Si_{20}O_{48}) \cdot 12H_2O$	Monoclinic	28.2
Other	Heulandite*	$Ca(Al_2Si_7O_{18}) \cdot 6H_2O$	Monoclinic	29.3
	Brewsterite	$Sr(Al_2Si_6O_{16}) \cdot 5H_2O$	Monoclinic	28.2
	Epistilbite	$Ca(Al_2Si_6O_{16}) \cdot 5H_2O$	Monoclinic	27.8
	Stilbite*	$Na_2Ca_4(Al_{10}Si_{26}O_{72}) \cdot 28H_2O$	Monoclinic	30.1
	Yugawaralite	$Ca_4(Al_7Si_{20}O_{54}) \cdot 14H_2O$	Monoclinic	29.2
	Laumontite	$Ca(Al_2Si_4O_{12}) \cdot 4H_2O$	Monoclinic	28.7
	Ferrierite	$Na_4Mg_2(OH)_2(Al_6Si_{30}O_{72}) \cdot 18H_2O$	Orthorhombic	28.2
	Paulingite	$(K,Ca)_{120}((Al,Si)_{580}O_{1160}) \cdot 690H_2O$	Cubic	37.5

*See separate article.

low as 35 parts per billion (ppb) can be achieved. This behavior exploits the high water affinity of zeolites coupled with the sieving properties of their molecularly sized pores. For example, K-A zeolite has pores of ∼0.3-nanometer diameter and so will admit only those molecules that are smaller (water, ammonia, and so forth).

The discrimination between molecules on the basis of their size as compared to the zeolite pore size is the basis of several extremely important separation processes in a procedure known as molecular sieving. Separations such as linear-form branched hydrocarbons for octane enhancement of fuels [using a zeolite in its calcium-exchanged form (Ca-A);5A], and *p*-xylene from *o*-, *m*-isomers and ethylbenzene (using large-pore zeolites such as faujasite), are operated on commercial scale. In addition to separations of molecules on the basis of their size, zeolites can also separtate like-sized molecules on the basis of their polarity. An important example is the passage of air through a bed of zeolites of low Si/Al ratio. Nitrogen is preferentially absorbed because of its higher polarity, and an oxygen-enriched effluent may be obtained for use in oxidation-combustion catalysis, medical applications, and so forth. The bed may be regenerated by brief evacuation, and this pressure-swing process provides an economical alternative to cryogenic separation of nitrogen and oxygen. *See* CHEMICAL SEPARATION TECHNIQUES; CRYOGENICS; MOLECULAR SIEVE.

Ion exchange. The major commodity use of zeolites takes advantage of their cation-exchange ability and leads to their use in low-phosphate detergents. Sodium cations of zeolite Na-A are exchanged into solution to replace divalent cations, for example, calcium present in hard water, thereby softening it. Mineral zeolites have found utility in agricultural and wastewater treatment applications, where they either ion-exchange harmful metal ions out of the stream being treated or absorb ammonia by reacting, as their acidic form, to produce the absorbed ammonium ion. *See* ION EXCHANGE.

Catalysis. The most important and expanding area of application for zeolites is as heterogeneous catalysts. By combining the propeties of excellent thermal stability (>800°C or 1500°F), a pore size of molecular dimensions, and the ready introduction of a wide assortment of cations via ion exchange, numerous very selective catalysts can be prepared. Much of the important catalysis relies on the generation of highly acidic sites within the zeolite pores. Thermal decomposition of the ammonium ion form of many zeolites leads to superacidic porous solids that can catalyze a wide range of organic molecule transformations via protonation and carbocation formation. More than 90% of catalytic cracking catalysts used in the United States are zeolite based and are used to break down heavier molecular components of crude oil to lighter fractions used in fuels and gasoline. *See* CRACKING; REACTIVE INTERMEDIATES.

Lanthanum-cerium-exchanged faujasties are 10,000 times more active than the amorphous aluminosilicate catalysts used previously. The synthetic zeolite ZSM-5 has been developed commercially for the catalytic conversion of methanol to gasoline via successive dehydration and polymerization of nascent methylene (CH_2) fragments. This step provides the last link in a chemical chain from coal or biomass to liquid fuels. ZSM-5 is also used for alkylation of toluene (with methanol) and for isomerization of mixed xylenes to produce the desired pure *p*-xylene isomer used in polyester manufacture. These latter chemistries rely on the shape-selective nature of the pore system of the zeolite structure. Since almost all of the reactive surface area and attendant acid sites of the zeolite are located within the internal pore structure (>99%), the catalytic chemistries occur within the pores and are, therefore, being performed in a very restricted space. This physical space constraint limits the outcome of the chemistry to those products that either can fit into the available space or can migrate from the site of their formation to the exterior of the zeolite crystalline and hence escape to be collected.

In addition to the acid-catalyzed chemistries, metal-catalyzed processes can be effected by using metal ion–exchanged zeolites where the zeolite imprints a shape or size selectivity onto the chemistry performed by the metal sites. Platinum (Pt), ruthenium (Ru), and nickel (Ni) hydrogenation and dehydrogenation-dehydrocyclization catalysts have been developed that exhibit useful selectivities toward organic molecules that are explained on the basis of the sieving effect of the zeolite host. Other developments involve the production of crystalline frameworks substituted with transition-metal ions such as titanium (Ti), chromium (Cr), manganese (Mn), and iron (Fe). The titanosilicates show excellent activity in selective oxidation transformations such as phenol to hydroquinone, using hydrogen peroxide as oxidant. Many completely nonsilicate zeolite frameworks have begun to appear as the field matures and synthetic techniques and understanding develop. Structures based on tin chalcogenides, aluminophosphates, and other tetrahedral-ion frameworks have become available. *See* HETEROGENEOUS CATALYSIS.

Finally, several nontraditional uses of zeolites have been introduced where the zeolite structure is viewed either as a nano-sized reaction vessel or as a nanocontainer for novel optical, electronic, or polymeric materials. The concept of the shape-selective pore structure of a zeolite as an inorganic replacement for the tertiary structure of an enzyme (zeozyme) has been explored, and examples demonstrating oxygen transport, electron transfer, photosynthesis, and selective oxygenation have been reported. Novel composites can be prepared where the pores of the zeolite are filled with a second phase of a polymer, semiconductor, metal, and so forth. Different physical properties (such as nonlinear optical behavior or superconductivity) become apparent because of the extremely small size domains of the included material imposed by the zeolite structure. Any physical property that depends on the collective properties of atoms or molecules (such as conductivity or optical behavior) becomes profoundly affected by being forced to reside in this nanocrystalline regime. *See* COMPOSITE MATERIAL; NANOCHEMISTRY; NANOSTRUCTURE.

Norman Herron

Bibliography. D. W. Breck, *Zeolite Molecular Sieves,* 1984; Ch. Baerlocher, W. M. Meier, and D. Olson (eds.), *Atlas of Zeolite Framework Types*, 5th ed., 2001; D. R. Corbin and N. Herron, Designing zeolite catalysts for size and shape selective reactions, *J. Mol. Catal.*, 86:343–369, January 1994; J. C. Jansen (ed.), *Advanced Zeolite Science and Applications*, 1994; H. Van Bekkum, E. M. Flanigen, and J. C. Jansen (eds.), *Introduction to Zeolite Science and Practice,* 2d ed., 2001.

Zero

In mathematics, the concept zero is used in two ways: as a number and as a value of a variable. The positional system of number notation, developed first by the Babylonians with the base 60, and later by the Hindus and the Chinese with the base 10, required for greater clarity a special marker of the empty, nonoccupied position. As such, the symbol for zero was introduced by the Babylonians (about 500 B.C.), and a millennium later it appeared in the Indian system of notation, which came to the West through the Arabs as the arabic number system.

The zero as a number, however, is a new concept, introduced by the Hindus and Chinese about the same time (sixth century). Brahmagupta (born A.D. 598) remarked that the number 0 has special properties: $a \pm 0 = a$, and $a \cdot 0 = 0$, where a may be any number (integer). Using zero as a denominator, he stated that $a/0 \pm b = a/0$, where $a/0$ thus appears as a number of a new kind, which cannot be changed by addition or subtraction. (In the modern way of thinking, division by zero is therefore not a "permissible" operation.) In the acceptance of zero as a number, the Orient is centuries ahead of the Occident. Leonardo Fibonacci (*Liber Abaci*, 1228) was the first Westerner to treat the zero. It is of interest that negative numbers appear earlier in the history of mathematics than does the number zero; they were already known to the Chinese mathematician Liu Hui in the third century of the Christian era. The historically late understanding of the number zero is witnessed by the fact that the Julian calendar (established 46 B.C.), from which the present-day Gregorian calendar is derived, has no year 0.

In a modern way, zero can be called the identity element of the infinite Abelian additive group of integers. If in an integral domain (all the more certainly in a field) a product is equal to zero, then at least one factor of the product is zero.

In the second concept zero is the value of a variable for which a function is equal to zero. For example, "A polynomial of degree n has n zeros," or "The

Riemann zeta function $\zeta(s)$ has all its complex zeros in the strip $0 <$ real part $s < 1$." *See* NUMBER THEORY.

Hans Rademacher; Emil Grosswald

Zinc

A chemical element, Zn, atomic number 30, and atomic weight 65.38. Zinc is a malleable, ductile, gray metal. Because of chemical similarities among zinc, cadmium, and mercury, these three metals are classed together in a transition-elements subgroup of the periodic table. *See* METAL; PERIODIC TABLE; TRANSITION ELEMENTS.

	1																18
	1 H	2										13	14	15	16	17	2 He
	3 Li	4 Be										5 B	6 C	7 N	8 O	9 F	10 Ne
	11 Na	12 Mg	3	4	5	6	7	8	9	10	11	12	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
	55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
	87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

Fifteen isotopes of zinc are known, of which five are stable, having atomic masses of 64, 66, 67, 68, and 70. About half of ordinary zinc occurs as the isotope of atomic mass 64. The half-lives of the radioactive isotopes range from 88 s for ^{61}Zn to 244 days for ^{65}Zn.

Zinc is a fairly active metal chemically. It can be ignited with some difficulty to give a blue-green flame in air and to discharge clouds of zinc oxide smoke. Zinc ranks above hydrogen in the electrochemical series, so that metallic zinc in an acidic solution will react to liberate hydrogen gas as the zinc passes into solution to form dipositively charged zinc ions, Zn^{2+}. This reaction is slow with very pure zinc, but the presence of small amounts of impurities, addition of a trace of copper sulfate, or contact between the zinc surface and such metals as nickel or platinum facilitates formation of gaseous hydrogen and speeds the reaction. The combination of zinc and dilute acid is often used to generate small quantities of hydrogen in the laboratory. Zinc also dissolves in strongly alkaline solutions, such as sodium hydroxide, to liberate hydrogen and form dinegatively charged tetrahydroxozincate ions, $Zn(OH)_4^{2-}$, sometimes written as ZnO_2^{2-} in the formulas of the zincate compounds. Zinc also dissolves in solutions of ammonia or ammonium salts. The common soluble zinc compounds undergo to some extent the process of hydrolysis, which makes their solutions slightly acidic. The ion Zn^{2+} is colorless, so that the relatively few zinc compounds that are not colorless in large crystals, or white as powders, receive their color through the influence of the other constituents. Some of the atomic and ionic properties of zinc are shown in the **table**. *See* ELECTROCHEMICAL SERIES; HYDROLYSIS.

Atomic and ionic properties of zinc

Property	Value
Electronic configuration	$1s^2, 2s^2, 2p^6, 3s^2, 3p^6, 3d^{10}, 4s^2$
Ionization potentials	
1st electron loss	9.39 eV
2d electron loss	17.9 eV
Ionic radius, Zn^{2+}	0.072 nm
Covalent radius (tetrahedral)	0.131 nm
Oxidation potentials	$Zn \rightleftharpoons Zn^{2+} + 2e^-$, $E^\circ = 0.76$ V $Zn + 4OH^- \rightleftharpoons ZnO_2^{2-} + 2H_2O + 2e^-$, $E^\circ = 1.22$ V

Zinc also forms many coordination compounds. The zincates are actually coordination compounds, or complexes, in which hydroxide ions, OH^-, are bound to the zinc ions. Ammonia, NH_3, forms complexes with zinc, such as the typical tetrammine zinc ion, $[Zn(NH_3)_4]^{2+}$. Zinc cyanide, usually given the simple formula $Zn(CN)_2$, is a coordination compound in which many alternating zinc and cyanide ions are three-dimensionally bound together in a very large molecule. This compound is still widely used in zinc plating, but concern over environmental pollution has led to increasing use of zinc chloride plating baths. In most coordination compounds of zinc, the fundamental structural unit is a central zinc ion surrounded by four coordinated groups arranged spatially at the corners of a regular tetrahedron. *See* COORDINATION CHEMISTRY; COORDINATION COMPLEXES.

Pure, freshly polished zinc is bluish-white, lustrous, and moderately hard (2.5 on Mohs scale). Moist air brings about a superficial tarnish to give the metal its usual grayish color. Pure zinc is malleable and ductile enough to be rolled or drawn, but small amounts of other metals present as contaminants may render it brittle. Malleability of even pure zinc is improved by heating zinc to 100–150°C (212–300°F). If heated zinc is mechanically worked; it does not embrittle on cooling. Zinc melts at 420°C (788°F) and boils at 907°C (1665°F). Its density is 7.13 times that of water, so that 1 ft^3 (0.028 m^3) of zinc weighs 445 lb (200 kg).

As a conductor of heat and of electricity, zinc ranks fairly high. However, its electrical resistivity (5.92 microhm-cm at 20°C or 68°F) is almost four times that of silver, the best conductor. As a conductor of heat, zinc is likewise only about one-fourth as efficient as silver. At 0.91 K zinc is an electrical superconductor. Pure zinc is not ferromagnetic, but the alloy compound $ZrZn_2$ displays ferromagnetism below 35 K.

The most important uses of zinc are in its alloys and as a protective coating on other metals. Coating iron or steel with zinc is known as galvanizing, and it may be done by immersing the article in melted zinc (hot-dip process), depositing zinc electrolytically onto the article in a plating bath (electrogalvanizing), exposing the article to powdered zinc near

its melting point (sherardizing), or spraying the article with melted zinc (metallizing). The mere physical presence of the zinc coat prevents corrosion of iron, and even if breaks in the coat expose portions of the iron, the greater chemical activity of the zinc causes it to be consumed in preference to the iron. Adding small amounts of other metals to galvanizing baths has been found to improve the adhesion and weathering qualities of the coating. *See* ELECTROPLATING OF METALS; GALVANIZING.

Even such nonstructural materials as cardboard can be zinc-coated by low-temperature flame spraying. Other important uses of zinc are in brass and zinc die-casting alloys, in zinc sheet and strip, in electrical dry cells, in making certain zinc compounds, and as a reducing agent in chemical preparations.

A so-called tumble-plating process coats small metal parts by applying zinc powder to them with an adhesive, then tumbling them with glass beads to roll out the powder into a continuous coat of zinc. Rechargeable nickel-zinc batteries offer higher energy densities than conventional dry cells. Foamed zinc metal has been suggested for use in lightweight structures such as aircraft and spacecraft. Some other uses of zinc are in dry cells, roofing, lithographic plates, fuses, organ pipes, and wire coatings. Zinc dust, a flammable material when dry, is used in fireworks and as a chemical catalyst and reducing agent. Radioactive ^{65}Zn is used medically in the study of metabolism of zinc, and also in determining rates of wear for zinc-containing alloys. *See* METAL COATINGS; ZINC ALLOYS.

William E. Cooley

Zinc is believed to be needed for normal growth and development of all living species, including humans; actually, life without zinc would be impossible. Zinc is a common element that is present in virtually every type of human food, and zinc deficiency is therefore not considered to be a common problem in humans. Zinc is a trace element; that is, it is present in biological fluids at a concentration below 1 ppm, and only a small amount (normally <25 mg) is required in the daily diet. (The recommended daily allowance for zinc is 15 mg/day for adults and 10 mg/day for growing children.) It is relatively nontoxic, without noticeable side effects at intake levels of up to 10 times the normal daily requirement.

James F. Riordan; Kenneth H. Falchuk

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; D. R. Lide, *CRC Handbook Chemistry and Physics*, 85th ed., CRC Press, 2004; W. Maret (ed.), *Zinc Biochemistry, Physiology and Homeostasis: Recent Insights and Current Trends*, 2002; S. W. Morgan, *Zinc and Its Alloys and Compounds*, 1985.

Zinc alloys

Combinations of zinc with one or more other metals. If zinc is the primary constituent of the alloy, it is a zinc-base alloy. Zinc also is commonly used in varying degrees as an alloying component with other base metals, such as copper, aluminum, and magnesium. A familiar example of the latter is the association of varying amounts of zinc (up to 45%) with copper to produce brass, the third largest use of zinc in the United States. *See* BRASS; COPPER ALLOYS.

Uses. Zinc-base alloys have two major uses: for casting and for wrought applications. Casting includes both die casting and gravity casting, which differs from die casting primarily in that no pressure is applied, except the force of gravity, in forcing the molten metal into the mold. *See* METAL CASTING.

Zinc alloys have historically been used for die casting mainly because of their relatively low melting temperatures. This means lower fabrication costs and longer tool life. Zinc-alloy die castings also have good strength and dimensional tolerance and exceptional castability and are easily plated and machined.

Alloys for casting. The zinc alloys used for die casting are well standardized and have been covered by American Society for Testing and Materials (ASTM) specifications in the United States since 1931. Equivalent specifications are in effect by similar organizations throughout the world, including the British Standards Institution (BSI), the Canadian Standards Association (CSA), and the International Standards Organization (ISO).

The presence of impurities above a certain specified maximum is deleterious in all types of alloys. Zinc die-casting alloys, like other alloys, are sensitive to small variations of certain impurities. The alloys, however, are not difficult to make, and most large die casters do their own alloying.

Lead, tin, and cadmium must be avoided as impurities in zinc, or as contaminants in the other alloying ingredients used. These impurities are high-density metals and have low melting points and, when present in amounts beyond the specification limits, render the alloys susceptible to intergranular corrosion, particularly on exposure to warm moist environments. To avoid these problems Special High-Grade Zinc, which is 99.99% pure, is used for die-casting applications. The great bulk of zinc die casting currently is carried out with two alloys, commonly identified as alloys 3 and 5, which have been in use for several decades. A third alloy was added and is commonly designated as alloy 7. The entire group in the United States is often spoken of by its original trade name as Zamak alloys, and in England as the Mazak alloys. A fourth alloy that is especially suited for cold-chamber die-casting machines was introduced to the industry in the mid-1960s. The composition and the mechanical and other properties of these alloys are given in **Table 1**, along with corresponding ASTM and Society of Automotive Engineers (SAE) designations.

Aluminum, the major alloying constituent in alloys 3 and 5, is added in amounts of about 4%, which has proved to be the optimum composition from the standpoint of strength, ductility, and stability. Loss of strength and stability may be as high as 50% of the original values if the aluminum content is reduced to 2%. The aluminum content also sharply reduces the rate of attack of molten zinc on iron containers. Aluminum also avoids "soldering," or sticking

TABLE 1. Composition and properties of zinc alloy die castings

ASTM designation B86-64:	AG40A	AC41A	–	–
SAE designation:	903	925	903	–
General designation:	3	5	ZAMAK 7	ILZRO 16
Composition % by weight				
Copper	.25 max	.75–1.25	.25 max	1.0–1.5
Aluminum	3.5–4.3	3.5–4.3	3.5–4.3	.01–.04
Magnesium	.020–.05	.03–.08	.005–.02	.01
Iron max.	.100	.100	.075	–
Lead max.	.005	.005	.0030	.004
Cadmium max.	.004	.004	.0020	.003
Tin max.	.003	.003	.0010	.002
Nickel	–	–	.005–.020	–
Titanium	–	–	–	.15–.25
Zinc (99.99 + % purity)	Remainder	Remainder	Remainder	Remainder
Mechanical properties				
Charpy impact strength, ft-lb (N·m), 1/4 × 1/4 in.				
(6.35 × 6.35 mm) bar, as cast	43 (58)	48 (65)	40 (54)	–
after 10 years of indoor aging	41 (56)	40 (54)	41 (56)	–
after 20 years of indoor aging	39 (53)	20 (27)	–	–
Tensile strength, lb/in.^{2} (MPa), after 20 years of indoor aging	33,000 (228)	36,000 (248)	–	–
Elongation, % in 2 in. (50.8 mm), after 20 years of indoor aging	20	12	–	–
Tensile strength, lb/in.^{2} (MPa), as cast	41,000 (283)	47,600 (328)	41,000 (283)	33,000 (228)
after 10 years of indoor aging	35,000 (241)	39,300 (271)	35,000 (241)	–
Elongation, % in 2 in. (50.8 mm), as cast	10	7	10	5–6
after 10 years of indoor aging	16	13	16	–
Expansion (growth), in. per in. (cm per cm), after 10 years of indoor aging	.0001	.0001	.0001	–
Other properties (as cast)				
Brinell hardness	82	91	76	75–77
Compression strength, lb/in.^{2} (MPa)	60,000 (414)	87,000 (600)	60,000 (414)	–
Electrical conductivity, mhos/cm^3 at 20°C	157,000	153,000	157,000	119,000
Melting point, °C	386.6	386.1	386.6	418
°F	727.9	727.0	727.9	785
Modulus of rupture, lb/in.^{2} (MPa)	95,000 (655)	105,000 (724)	95,000 (655)	–
Shear strength, lb/in.^{2} (MPa)	31,000 (214)	38,000 (262)	31,000 (214)	–
Solidification point, °C	380.6	380.4	380.6	416
°F	717.1	716.7	717.1	780
Solidification shrinkage, in./ft (cm/m)	.14 (1.2)	.14 (1.2)	.14 (1.2)	–
Specific gravity	6.6	6.7	6.6	7.1
Specific heat, cal/g/°C [J/(g·°C)]	0.10 (0.42)	0.10 (0.42)	0.10 (0.42)	–
Thermal conductivity, cal/(s·cm·°C) [J/(s·cm·°C)] at 18°C	0.27 (1.13)	0.26 (1.09)	0.27 (1.13)	0.205 (0.86)
Transverse deflection, in.	0.27	0.16	0.27	–
Thermal expansion per °C	.0000274	.0000274	.0000274	.000027
per °F	.0000152	.0000152	.0000152	.000015
Specific weight, lb/in.^{3} (g·cm^3)	.24 (6.6)	.24 (6.6)	.24 (6.6)	.26 (7.2)

of the casting to the die. This permits the casting of these zinc alloys in the more productive hot-chamber (plunger) type of die-casting machine. This type of machine injects the molten metal into the die cavity automatically, unlike the cold-chamber machine which requires the molten metal to be poured, by ladling, into the cylinder. *See* ALUMINUM.

Alloys 3 and 5 differ primarily in their copper content. Copper additions increase strength and hardness and improve corrosion resistance. The copper content of alloys is held to a specified maximum of 1.25% (in alloy 5) to avoid making the alloy unstable through aging and to avoid reducing the impact strength to a very low value. Dimensional stability also is affected by high levels of copper, so that a 1% maximum is usually recommended. Changes in casting dimensions, however, are minimal and always occur within a few weeks after casting; consequently, unless the castings are large or require very close tolerances, the dimensional changes can usually be considered insignificant. In those cases where very high dimensional accuracy is required, alloy 3 is usually recommended.

Alloy 7 is considered a modification of alloy 3. It has a lower magnesium content, and iron, lead, tin, and cadmium are also held to lower levels. A small amount of nickel is added. When the alloys are compared, alloy 7 exhibits improved casting properties, making it easier to secure high-quality surfaces for finishing (mostly chromium plating) and to obtain higher production rates. The improved castability is due primarily to the low level of magnesium.

The alloy ILZRO 16 has a fairly wide permissible composition range, but in practice it is recommended that it be taken as titanium 0.15–0.25%, chromium 0.01–0.20% (Ti plus Cr should range between 0.30 and 0.40%), copper 1.0–1.5%, and aluminum 0.10–0.04%, the remainder being zinc. This alloy has a superior creep resistance and is designed for applications required to resist deformation under a sustained stress, especially at elevated temperatures. A comparison of the creep properties of ILZRO

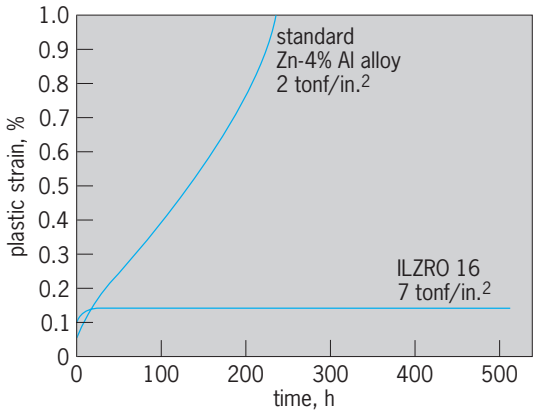


Fig. 1. Creep of two pressure-die-cast zinc alloys at 167°F (75°C). 1 ton-force/in.^{2} = 13.8 MPa.

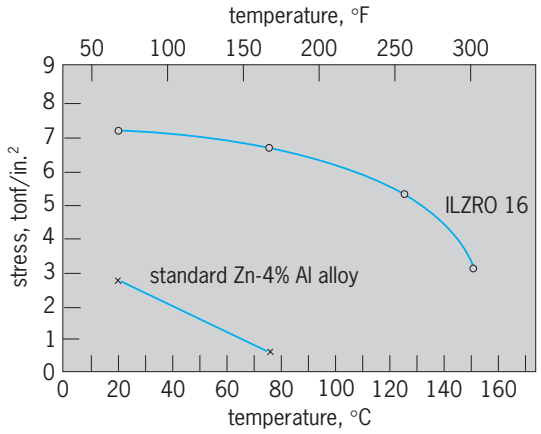


Fig. 2. Variation with temperature and stress to produce 0.1% plastic strain in 500 h for two zinc die-casting alloys. 1 tonf/in.^{2} = 13.8 MPa.

16 and the conventional 4% aluminum alloy is shown in **Fig. 1**. The effect of temperature on the stress that is required to produce 0.1% plastic strain in 500 h is shown in **Fig. 2**.

An important gravity-casting alloy, known as ILZRO 12, was introduced to the industry in 1967. This foundry alloy can be cast in sand, plaster, or permanent molds, is economic for general foundary work, and has found considerable application for prototypes. Its nominal composition is 10.5–11.5% aluminum, 0.5–1.25% copper, 0.01–0.03% tin (maximum), and balance zinc by difference. Principal properties of ILZRO 12 are shown in **Table 2**. There is some improvement in impact resistance as the aluminum content is reduced. Reducing the content to around 10% improves ductility, but tensile strength is lower and impact is about the same as in 11% aluminum alloys.

This new foundry alloy has high tensile strength, excellent castability, and other advantages inherent in zinc die-casting alloys. One of its most important characteristics is that its as-cast properties closely relate to those of the conventional zinc alloys when die cast. The rate of cooling has no apparent effect on its mechanical properties. The following are among its other advantages: foundry scrap can be recycled repeatedly; gating of dies is simple; it is not sensitive to superheating, stewing, or remelting; it has low shrinkage; and gas porosity is minimized. In addition, it can be plated or otherwise finished using conventional finishing processes.

Wrought zinc alloys. In the wrought zinc area, numerous compositions and alloys are used, depending on ultimate product requirements. Alloying metals can be used to improve various properties, such as stiffness, for special applications. The zinc-copper-titanium alloy has become the dominant wrought-zinc alloy for applications demanding superior performance. Composition of this alloy is 0.4–0.8% copper, 0.08–0.16% titanium, 0.3% max. lead, 0.015% max. iron, 0.01% max. manganese, 0.01% max. cadmium, 0.02% max. chromium, and the balance Special High-Grade Zinc.

The conventional grades of slab or ingot zinc are

TABLE 2. Properties of ILZRO 12

Property	Value
Yield strength, lb/in.^{2} (MPa)	
Sand cast	30,000 (207)
Chill cast	31,000 (214)
Tensile strength, lb/in.^{2} (MPa)	
Sand cast	43,500 (300)
Chill cast	53,500 (369)
Impact strength, ft-lb (N·m)	20 (27)
Shear strength, lb/in.^{2} (MPa)	32,250 (222)
Elongation, % in 2 in. (50.8 mm)	
Sand cast	3 1/2
Chill cast	5
Hardness (Brinell–500 kg)	
Sand cast	110
Chill cast	115
Compressive 0.1% proof stress, lb/in.^{2} (MPa)	21,000 (145.5)
Creep strength to give 0.1% strain at 500 h, lb/in.^{2} (MPa)	
68°F (20°C)	8000* (55)
122°F (50°C)	3000 (21)
Fatigue endurance limit, 10^8 cycles, lb/in.^{2} (MPa)	8000 (55)
Properties after 200 days at 212°F (100°C)	
Tensile strength, lb/in.^{2} (MPa)	30,130 (207.7)
0.1% proof stress, lb/in.^{2} (MPa)	17,800 (122.7)
0.2% proof stress, lb/in.^{2} (MPa)	21,160 (145.9)
Elongation, % in 2 in. (50.8 mm)	6
Impact strength, ft-lb (N·m)	2 1/2 (3.4)
Specific gravity	6.0
Density, lb/in.^{3} (g/cm^3)	0.216 (6.0)
Steam test	Resistant
Dimensional stability at 68°F (20°C)	0.0001 in./in./year (0.001 cm/cm/year) Shrinkage during first year
Thermal expansion at	
68–212°F	1.55×10^{-5} per °F
20–100°C	2.79×10^{-5} per °C
68–392°F	1.61×10^{-5} per °F
20–200°C	2.90×10^{-5} per °C
Thermal conductivity	0.21–0.22 cal/(s · cm · °C) 0.88–0.92 J/(cm · °C · s)

*Similar to zinc–4% aluminum pressure-die-casting alloy.

rolled into sheets, strips, ribbons, foils, plates, or rods for a wide variety of uses. The zinc alloy also may be continuous-cast into rod or bar. Because of its properties, rolled zinc can be easily worked into various shapes and forms by common fabricating methods, including stamping, forming, and spinning. It can be polished and lacquered to retain its natural bright luster, or it can be plated or painted for other finishing effects. It can also be left to weather naturally, in which case it forms an attractive, nonstaining patina, making it especially suited for architectural applications.

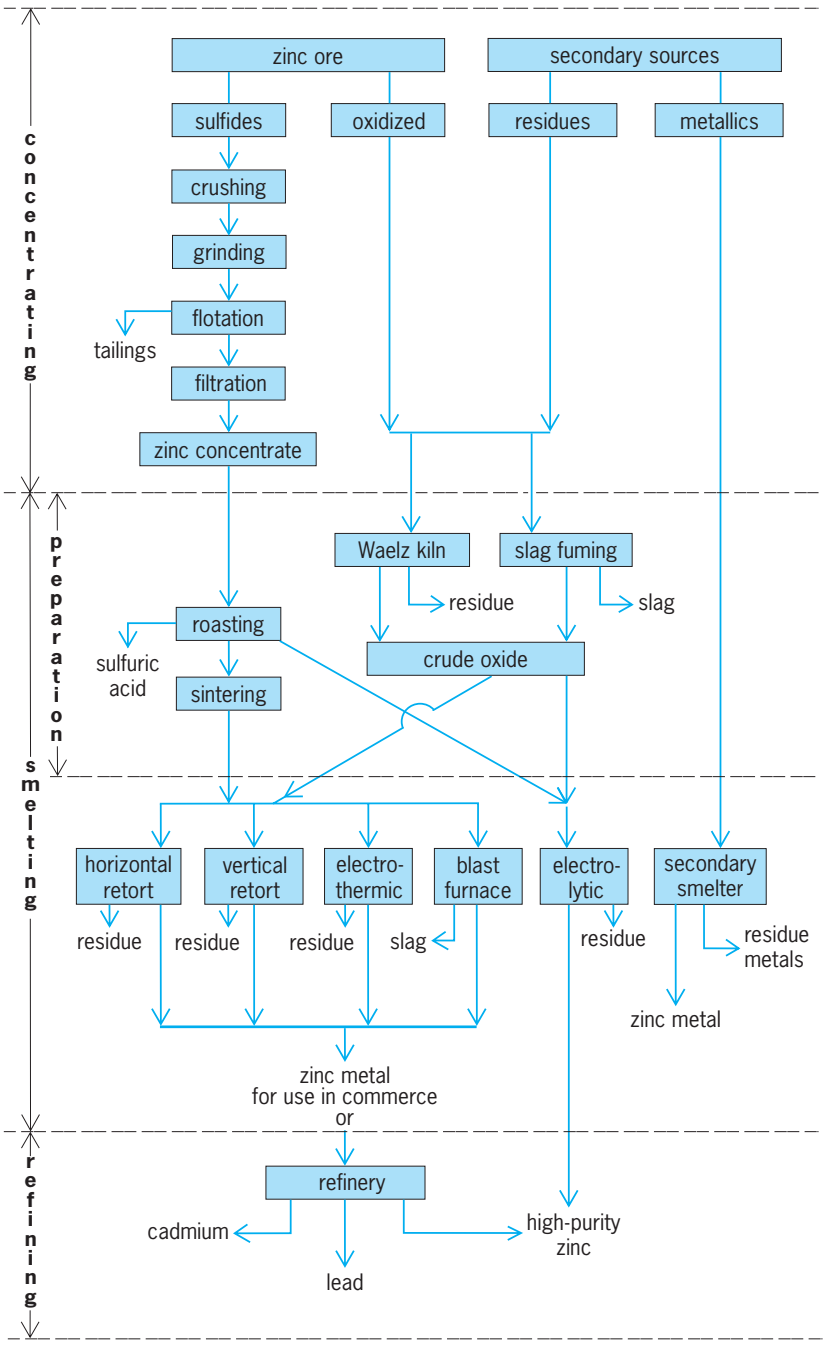
The zinc-copper-titanium wrought-zinc alloy has all of these properties and on a comparative basis is stronger and more dent-resistant than some other metals of the same thickness. Tensile strength is 29,000 lb/in.^{2} (200 megapascals) with the grain and 40,000 lb/in.^{2} (280 MPa) across the grain. Elongation over 2 in. (5 cm) is 26% with the grain and 14% across the grain. Rockwell hardness is 60–76. It is easily soldered, is nonmagnetic, and has an electrical conductivity equal to that of brass. It is produced in coils ranging 0.003–0.250 in. (0.008–0.635 cm) thick and 0.25–19 in. (0.635–48 cm) wide, and in sheets ranging 0.020–0.125 in. (0.051–0.318 cm) thick, 20–60 in. (51–153 cm) wide, and to 120 in. (305 cm) long. *See* ZINC; ZINC METALLURGY. Adolph L. Ponikvar

Bibliography. S. Gupta, *The World Zinc Industry*, 1981; C. H. Matthewson (ed.), *Zinc: The Science and Technology of the Metal, Its Alloys and Compounds* ACS Monogr. 142, 1959; S. W. Morgan, *Zinc and Its Alloys*, 1977.

Zinc metallurgy

Separating and extracting zinc from ores, refining it, and preparing it into usable forms. Ordinarily, after being mined, ores must first be separated into a concentrated mineral and a waste rock. This concentrate then is reduced to the metal in a metallurgical works. Finally, the metal may be further refined and alloyed to commercially usable form. Commonly, these three extractive metallurgical operations are conducted at separate locations and are broadly categorized as concentrating, smelting, and refining, respectively. At each operation there are several process steps necessary to accomplish the overall purpose. The **illustration** shows the various interrelated processes of zinc metallurgy. *See* METALLURGY; ZINC.

Zinc-bearing ores of commercial value occur predominantly in sulfide form (for example, sphalerite, ZnS). Some minor sources are known as oxidized ores and occur as zinc oxide, zinc silicate, zinc carbonate, and some complex combinations of these and other elements. Usual impurities in zinc ores will be one or more of the following: lead, cadmium, and iron; frequently copper, silver, gold, arsenic, and antimony; and occasionally tin, indium, germanium, and gallium. The primary objective of extractive metallurgical operations, then, is to separate the zinc from these impurities. Of major importance is the by-product recovery of some of these impurities



Schematic flow sheet for zinc metallurgy.

in economic form during the processing for zinc extraction. *See* SPHALERITE.

Concentration. Zinc sulfide ores are usually beneficiated (concentrated or ore-dressed) adjacent to the mine site. First, the ore must be crushed and ground in order to free the mineral lattices from those of the waste rock (gangue). Next, the finely divided ore is mixed into a slurry with water and the mineral and gangue particles are separated utilizing the effect of gravity. The most common means of separation is the froth flotation process, wherein a small amount of pine oil is added to the slurry, followed by air agitation. A thin oil film entraps a small bubble of air which is then attracted to the fine mineral sulfide

particle. As they float to the top of the vessel, these bubbles (froth) carrying the mineral are skimmed from the overflow and collected.

The gangue does not attract the bubble and is removed from the bottom of the vessel to be discarded as waste (tailings). Further selective (or differential) flotation steps may be performed, using suitable chemical reagents, to separate much of the lead sulfide and iron sulfide minerals from the zinc sulfide, in producing a purer zinc concentrate. The final step in beneficiation consists of merely separating the mineral slurry into solids and water by means of a filtration process. Resulting zinc sulfide concentrates typically contain 50–64% zinc, depending upon the extent of other impurities which remain to be eliminated in ensuing metallurgical processing operations. Oxidized zinc ores and other zinc-bearing secondary materials are generally not susceptible to concentration at their source, this function usually being performed in conjunction with the smelting operation. *See* ORE DRESSING.

Preparation for smelting. This second operation of zinc metallurgy involves first the chemical or physical changes required in the source material to prepare it into a crude zinc oxide form and a specified particle size (depending upon the particular smelting method contemplated). Such process steps variously entail the operations of roasting, sintering, or pyroconcentration.

Roasting. Zinc sulfide concentrates must be converted to the oxide form prior to smelting. This is accomplished by heating the sulfide and burning it with oxygen (usually from air) to form crude zinc oxide and gaseous sulfur dioxide. The latter is usually converted to sulfuric acid with suitable equipment, and it yields an economic credit for the roasting operation. Production of sulfuric acid is mandatory in many locations to abate an air-pollution problem which would otherwise result.

Several types of roasting plants are in satisfactory operation, including rabbled single-hearth and multiple-hearth furnaces; suspension (flash) roasters in which a showering stream of finely divided sulfide concentrate is burned in a combustion chamber; and fluidized-bed roasters, in which the combustion air is pressurized, being admitted through very small tuyere holes in a steel plate forming the base of the combustion chamber and burning the sulfur from the sulfide concentrates as the gas stream flows upward through the hot bed. Most modern zinc roasters are of the fluid-bed type for reasons of economy and metallurgical efficiency.

Sintering. Most pyrometallurgical extraction methods require that the roasted or pyroconcentrated oxides (and sometimes a portion of concentrates) be further treated to increase their density and particle size prior to feeding into the smelter. Suitable control of temperature, airflow, and reducing agents in this step also serves to remove impurities from the crude zinc oxide during sintering. These are driven off chiefly as volatilized compounds of cadmium, lead, silver, and indium which are carried over in the combustion gas stream and captured in a bag house or electrostatic precipitator for subsequent recovery of these metals. *See* SINTERING.

Equipment for performing the sintering operation is usually a Dwight-Lloyd moving grate machine or a rotary kiln.

Pyroconcentration. Oxidic zinc ores of low grade, secondary zinc-bearing materials, and certain residues from other chemical and metallurgical operations are often concentrated economically in a rotary kiln (Waelz) operation or alternatively by means of the so-called slag-fuming process. In either case, the zinc-bearing material is mixed with sufficient coal to reduce the oxide zinc to its elemental form. Heat required for this endothermic chemical reaction and for vaporization of the zinc is provided partially by burning of natural gas or powdered coal inside the chamber and partially by oxidation of the carbon monoxide and zinc vapor by excess air in the traveling gas stream. Resultant zinc oxide (of much higher concentration than in the feed) is carried away from the reaction zone in the gas stream and is thence collected in a bag-house (filter) and forwarded to the smelter. Nonvolatile impurities, such as iron, copper, and silica, remain in the slag residue and are usually discarded as waste.

Smelting methods. The other basic part of smelting is the reduction step, wherein the zinc is reduced from its oxide to its elemental form. Several very successful and varied processes accomplish this function, including horizontal retort, vertical retort, electrothermic furnace, and blast furnace (all of which are pyrometallurgical and use carbon as a reducing agent). The electrolytic process uses the passage of electric current for reduction to metal from a liquid bath (hydrometallurgical). Used zinc is also reclaimed in secondary smelters. In all pyrometallurgical processes the basis for this step is the requirement for sufficient heat generation to raise the temperature of the zinciferous material and reduction fuel (usually coal and coke mixture) to at least 1832°F (1000°C), so that the carbon will reduce the zinc oxide to zinc metal. This is above the boiling point of zinc, thus it is in the vapor phase. Carbon above that required for zinc reduction must be present to shift the equilibrium away from formation of carbon dioxide and toward formation of carbon monoxide; otherwise the former will reoxidize the zinc vapor as it passes out of the reaction chamber into the condensation vessel. *See* PYROMETALLURGY, NONFERROUS.

Most impurities remaining from the preparation steps described in the previous topic are eliminated to the furnace residues, including cadmium and much of the lead.

The hydrometallurgical (electrolytic) process for zinc smelting, rather than using heat for the reduction, relies on electrodeposition of the metal from a zinc sulfate solution prepared from the crude zinc oxide and sulfuric acid. Virtually all impurities remaining from the preparation step are eliminated in this process. *See* HYDROMETALLURGY.

Horizontal (Belgian) retort process. This process is the earliest known method of zinc smelting and employs vast, honeycomblike batteries of fire clay or silicon-carbide retorts (800 or so in a block), 8–10 in. (20–25 cm) in diameter and 5–6 ft (1.5–1.8 m) long, set

horizontally in a gas- or coal-fired furnace. Retorts are supported at each end and set in rows of 15–20, 4–5 high. The mixed feed of sintered concentrate and coal is usually charged by mechanical means into one set of retorts at a time. As the reaction temperature is attained, a carbon monoxide gas and zinc vapor mixture issues from the mouth of the retort, the zinc being condensed in a clay vessel which has been luted to the open retort end. The mouth of this condenser is stuffed with a porous coke mixture, allowing the noncondensable carbon monoxide to pass through and burn while the molten zinc is collected inside. Four or five times during the 24–48-h firing cycle, zinc is tapped from these individual condensers into a transfer ladle, thence poured into ingot molds.

At cycle end, the condensers are taken down from an entire furnace and the residues cleaned out by machines to make the retorts ready for receiving new charge. Development of charging and cleanout machines has significantly reduced physical labor required in this process. Further labor-saving efforts have been made to centralize condensation and metal tapping by replacing the hundreds of individual condensers on a furnace with an airtight metal shield across the face of the furnace, but conclusive results have not yet been established. The number of operating horizontal retort zinc smelters has declined markedly due to the difficulty of preventing particulate emissions, the stringencies in natural gas supply, and the high labor requirements.

Vertical (New Jersey) retort process. In this process a vertical retort of silicon carbide brick, having a horizontal rectangular dimension of 1 ft (0.3 m) by 6–8 ft (1.8–2.4 m) and a vertical heated height of up to 35 ft (10.7 m), is employed. Gas heats the external walls and a briquetted charge of mixed roasted concentrate and coal is put through a coking process, then admitted through the top at frequent intervals. Residue discharges continuously from the bottom. The mixed zinc vapor–carbon monoxide gas stream is constantly drawn off near the top of the retort and the molten zinc collected in a special condenser, while the carbon monoxide is withdrawn for use as fuel in the external heating. Typical daily zinc production from one such vertical retort may be 10 tons (9 metric tons) whereas that for one entire block (of 800) horizontal retorts might be only 20 tons (18 metric tons).

Electrothermic (St. Joseph) retort process. An 8-ft-diameter (2.4-m), upright firebrick retort of up to 50 ft (15 m) high is the furnace in this process. Sintered concentrate and coke mixture is charged constantly into the top of the furnace where heat of reaction is supplied by passage of heavy electrical current through the resistance of the charge. Sixteen large-diameter graphite electrodes, eight near the top and eight near the bottom, transmit the current to the descending charge. Residue is continuously withdrawn from the bottom. The mixture of zinc vapor and gas is drawn through a bath and condensed into metal, or it may be alternately burned into zinc oxide and collected in a bag house for use in the rubber or paint industries. A single furnace produces 80 tons (72 metric tons) or more of zinc per day.

Blast furnace (Imperial Smelting) process. Design of this furnace is taken from the process ordinarily used in smelting of lead. In fact, the ISP furnace smelts zinc and lead concomitantly. The firebrick vertical shaft is of a rectangular cross section, pinched in at the bottom to accommodate several tuyeres for injection of blast air. Heat of reaction is generated internally by burning the carbon content of the mixed charge. The latter is admitted in frequent batches at the top of the shaft and consists of sintered lead, zinc concentrate, and suitable fluxes, to produce a slag which is tapped at shaft bottom. Certain impurities are carried out as waste in the slag; valuable lead bullion and copper-bearing matte are separated from the slag at this stage. Carbon monoxide gas carrying the zinc vapor is drawn through a spray of molten lead, where the zinc is shock-cooled and conveyed by the molten stream to settling basins. Zinc, being less dense than lead, is withdrawn from the top of the bath and cast. Because of the unique method of collection, zinc produced from this furnace carries high lead content.

Electrolytic process. An increasing percentage of the world's zinc metal is produced by this method. In contrast with the pyrometallurgical processes, the electrolytic reduction of zinc concentrate is leached with sulfuric acid (largely regenerated as spent electrolyte in the cells).

The resulting zinc sulfate solution is rigorously purified in two or more stages before being pumped into the electrolytic cells. Lead-silver alloy anodes and aluminum cathodes are used, the pure zinc deposited on the latter being stripped periodically, melted, and cast into ingots. Virtually all impurities entering with the zinc concentrate are removed as residues in the closely controlled roasting, leaching, and purification steps, resulting in production of very high-purity zinc, directly. *See* ELECTROMETALLURGY.

Secondary smelting. Of increasing importance is the reclaiming of used zinc contained in chemical and metallurgical residues and from scrappage of appliances, automobiles, and aircraft dies. The nonmetallic forms are treated in primary and secondary smelters much as are oxidized ores, requiring carbon reduction. Metallic forms are frequently melted in muffle furnaces or special retorts, and the zinc is selectively vaporized from higher-boiling impurities, being suitably condensed and cast for reuse.

Usually, the metal reclaimed by the secondary smelting process will not meet specifications for the higher grades of zinc; however, careful preselection of the scrap to be melted may result in production of up to the High-Grade purity.

Refining. Purity of zinc produced by the various carbon reduction processes is largely dependent upon controls and operating procedures practiced during the smelting preparation steps. The quality produced is variously suitable for hot-dip galvanizing, continuous-line galvanizing, and in some cases for brass manufacture and rolled (wrought) zinc; however, for the sizable usage in die-casting alloys, output from this type of smelter must

undergo a refining step. Fractional distillation in reflux refining columns is the leading method of upgrading the lower-purity zinc metal.

Typically, a three-column unit is operated with two lead columns followed by a cadmium column. The lead column concentrates lead, iron, and other high-boiling impurities in the lower end, driving zinc, cadmium, and other lower-boiling metals over the top and into the cadmium column. Cadmium is then condensed off the top of this column while purified zinc is withdrawn from the bottom. The process is capable of producing zinc that is 99.995% pure.

Five commercial grades of zinc are recognized by the American Society for Testing and Materials, ranging downward in purity from Special High-Grade (99.99% minimum), through High-Grade, Intermediate, Brass Special, to the lowest grade, Prime Western (98.3% minimum), with maximum impurity limits being placed on each grade with respect to content of lead, iron, cadmium, aluminum, tin, or all of these. *See* METAL AND MINERAL PROCESSING; ZINC ALLOYS.

Carl H. Cotterill

Bibliography. S. Gupta, *The World Zinc Industry*, 1981; C. H. Matthewson (ed.), *Zinc: The Science and Technology of the Metal, Its Alloys and Compounds*, ACS Monogr. 142, 1959; S. W. Morgan, *Zinc and Its Alloys*, 1977.

Zincite

A mineral with composition ZnO (zinc oxide). It crystallizes in the hexagonal system with a wurtzite-type structure. Thus its principal axis is polar and different forms appear at top and bottom of crystals. Such crystals are rare and the mineral is usually massive. Cleavage is prismatic. Its hardness is 4 and its specific gravity 5.6. The mineral has a subadamantine luster and a deep-red to orange-yellow color. The color of the mineral is believed to result partly from the presence of manganese and partly from defects in the crystal structure. Zincite is rare except at the zinc deposits at Franklin and Sterling Hill, New Jersey. There, associated with franklinite and willemite, it is mined as a valuable ore of zinc. *See* ZINC.

Cornelius S. Hurlbut, Jr.

Zingiberales

An order of flowering plants, division Magnoliophyta, in the monocots, consisting of eight families and about 1800 species. Zingiberales (also known as Scitamineae or Scitaminales) are morphologically well defined and clearly circumscribed in DNA sequence analyses. The largest families are Zingiberaceae (about 1000 species), Marantaceae (about 400 species), Costaceae (about 150 species), and Heliconiaceae (about 100 species). Zingiberales are most closely related to Commelinales, and the enigmatic *Hanguana* from southeast Asia is difficult to place with certainty in either order. Zingiberales are herbs or scarcely branched trees or shrubs with pinnately veined leaves and irregular flowers that have well-differentiated sepals and petals, an inferior ovary, septal or septa-derived nectaries, and usually either one or five functional stamens. The endosperm is starchy, and the stomates have two or (usually) more subsidiary cells. The economic crops, ginger (*Zingiber officinale*) in Zingiberaceae and banana (*Musa*) in Musaceae, and the ornamentals, bird of paradise flower (*Strelitzia*) in Strelitziaceae and *Canna* in Cannaceae, are familiar members of Zingiberales. *See* BANANA; COMMELINALES; FLOWER; FRUIT; GINGER; LILIOPSIDA; MAGNOLIOPHYTA.

Michael F. Fay

Zingiberidae

A subclass of Liliopsida (Monocotyledons) of the division Magnoliophyta (Angiospermae), the flowering plants, containing two orders (Bromeliales and Zingiberales), with nine families and about 3800 species. The subclass has been associated with the Commelinidae and the Liliidae, sharing some features with each but aberrant in either, and it differs from both in usually having four or more supporting cells around each stomate. The flowers are usually epigynous, usually bisexual, and often irregular, and they are usually showy and adapted to pollination by insects or other animals. The pistil consists of three united carpels, often with septal nectaries opening at the top of the ovary. The seeds usually have well-developed, mealy or starchy endosperm. *See* BROMELIALES; LILIOPSIDA; MAGNOLIOPHYTA; PLANT KINGDOM; SECRETORY STRUCTURES (PLANT); ZINGIBERALES.

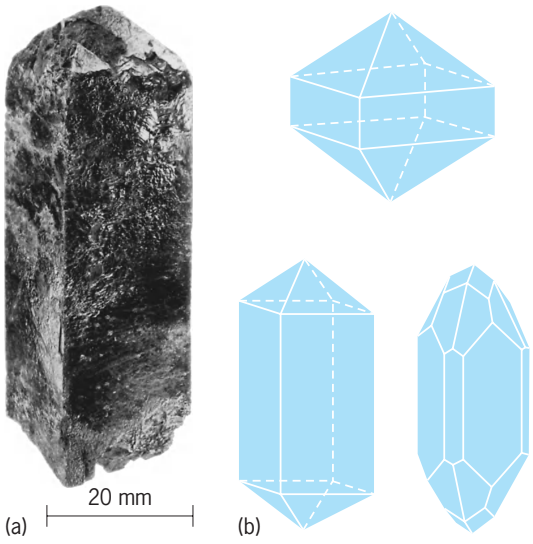
T. M. Barkley

Zircon

A mineral with the idealized composition $ZrSiO_4$, one of the chief sources of the element zirconium. Trace amounts of uranium and of thorium are often present and the mineral may then be partly or entirely metamict. The name cyrtolite is applied to an altered type of zircon. Structurally, zircon is a nesosilicate, with isolated SiO_4 groups. It is isostructural with the thorium silicate thorite and the yttrium phosphate xenotime. *See* RADIOACTIVE MINERALS; SILICATE MINERALS; ZIRCONIUM.

Zircon is tetragonal in crystallization. It often occurs as well-formed crystals, which commonly are square prisms terminated by a low pyramid (see **illus.**). The color is variable, usually brown to reddish brown, but also colorless, pale yellowish, green, or blue. The transparent colorless or tinted varieties are popular gemstones. Hardness is $7\frac{1}{2}$ on Mohs scale; specific gravity is 4.7, decreasing in metamict types. *See* METAMICT STATE.

Because of its chemical and physical stability, zircon resists weathering and accumulates in residual deposits and in beach and river sands, from which it has been obtained commercially in Florida and in India, Brazil, and other countries. It is widespread



Zircon. (*a*) Crystal from Eganville, Ontario, Canada (*American Museum of Natural History specimen*). (*b*) Crystal habits (*after C. Klein and C. S. Hurlbut, Jr., Manual of Mineralogy, 21st ed., John Wiley and Sons, 1993*).

in small amounts as an accessory mineral in granitic and syenitic igneous rocks, and occurs more abundantly in pegmatite deposits associated with such rocks. Zircon also is a minor constituent in many types of gneissic and schistose metamorphic rocks. *See* HEAVY MINERALS.

Clifford Frondel

Zirconium

A chemical element, Zr, atomic number 40, atomic weight 91.22. Its naturally occurring isotopes are 90, 91, 92, 94, and 96. Zirconium is one of the more abundant elements, and is widely distributed in the Earth's crust. Being very reactive chemically, it is found only in the combined state. Under most conditions, it bonds with oxygen in preference to any other element, and it occurs in the Earth's crust only as the oxide, ZrO_2, baddeleyite, or as part of a complex of oxides as in zircon, elpidite, and eudialyte. Zircon is commercially the most important ore. Zirconium and hafnium are practically indistinguishable in chemical properties, and occur only together. *See* HAFNIUM; PERIODIC TABLE; ZIRCON.

1												13	14	15	16	17	18
1 H	2											5 B	6 C	7 N	8 O	9 F	2 He
3 Li	4 Be																10 Ne
11 Na	12 Mg	3	4	5	6	7	8	9	10	11	12	13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112	113					

lanthanide series	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
actinide series	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

Most of the zirconium used has been as compounds for the ceramic industry: refractories, glazes, enamels, foundry mold and core washes, abrasive grits, and components of electrical ceramics, The incorporation of zirconium oxide in glass significantly increases its resistance to alkali. The use of zirconium metal is almost entirely for cladding uranium fuel elements for nuclear power plants. Another significant use has been in photo flashbulbs.

Zirconium is a lustrous, silvery metal, with a density of 6.5 g/cm^3 (3.8 oz/in.^{3}) at 20°C (68°F). It melts at about 1850°C (3362°F). Estimates of the boiling point from appropriate data have commonly been of the order of 3600°C (6500°F), but observations suggest about 8600°C (15,500°F). The free energies of formation of its compounds indicate that zirconium should react with any nonmetal, other than the inert gases, at ordinary temperatures. In practice, the metal is found to be nonreactive near room temperature because of an invisible, impervious oxide film on its surface. The film renders the metal passive, and it remains bright and shiny in ordinary air indefinitely. At elevated temperatures it is very reactive to the non-metallic elements and many of the metallic elements, forming either solid solutions or compounds.

Zirconium generally has normal covalency of 4, and commonly exhibits coordinate covalencies of 5, 6, 7, and 8. Zirconium is at oxidation number 4 in nearly all of its compounds, Halides in which its oxidation numbers are 3 and 2 have been prepared. While zirconium is often part of cationic or anionic complexes, there is no definite evidence for a monatomic zirconium ion in any of its compounds.

Most handling and testing of zirconium compounds have indicated no toxicity. There has generally been no ill consequence of contact of zirconium compounds with the unabraded skin. However, some individuals appear to have allergic sensitivity to zirconium compounds, characteristically manifested by appearance of nonmalignant granulomas. Inhalation of sprays containing some zirconium compounds and of metallic zirconium dusts have had inflammatory effects.

Warren B. Blumenthal

Bibliography. F. A. Cotton et al., *Advanced Inorganic Chemistry*, 6th ed., Wiley-Interscience, 1999; J. Hala (ed.), *Halides, Oxyhalides and Salts of Halogen Complexes of Titanium, Zirconium, Hafnium, Vanadium, Niobium and Tantalum*, 1989; M. F. Lappert, *Comprehensive Organometallic Chemistry II: Scandium, Yttrium, Lanthanides and Actinides, and Titanium, Zirconium, and Hafnium*, vol. 4, 1995.

Zoanthidea

An order of the cnidarian subclass Hexacorallia. Sometimes known popularly as "mat anemones" and technically also as Zoanthiniaria (the name Zoantharia is an alternative for subclass Hexacorallia), this order contains relatively few species. *See* ANTHOZOA; CNIDARIA; HEXACORALLIA.

Morphology. The column of one of these sedentary, anemonelike anthozoans is typically encrusted with sand grains, sponge spicules, tests of formaniferans, and other detritus that invades the mesoglea, thus acting as a sort of adventitious skeleton. Zoanthid musculature is poorly developed. The basal end of most is somewhat or deeply embedded in a mass of tissue that connects members of the colony; in solitary zoanthids, a pedal disk is absent. The simple, unbranched tentacles are arrayed in two cycles. As in some other hexacorallians, the mesenteries are paired and coupled, but otherwise the mesenterial arrangement is distinctive, differing slightly between two groups of zoanthids. In both groups—suborders Brachycnemina and Macrocnemina—the mesenteries that connect with the single siphonoglyph (termed the sulcal mesenteries, conventionally considered ventral in position) are complete, and the asulcal pair (those diametrically opposite the siphonoglyph, on the "dorsal" side of the animal) are incomplete. However, in the suborder Brachycnemina each nondirective pair (that is, each pair not connected to the siphonoglyph) consists of one complete and one incomplete mesentery (thus an animal has only one pair of complete mesenteries), whereas in the suborder Macrocnemina, both members of the second pair of mesenteries on each side of the asulcal pair are complete (thus an animal has three pairs of complete mesenteries). Mesenteries are added only in the exocoels on either side of the sulcal mesenteries; the larger mesentery of each pair bears filaments and is gametogenic.

Distribution. Members of Brachycnemina are abundant in warm, shallow waters, including on coral reefs, and typically harbor zooxanthellae; they comprise the families Neozoanthidae, Sphenopidae, and Zoanthidae. The Macrocnemina is characteristic of the deep sea, where colonies of some species encrust objects such as sea whips, and others associate with hermit crabs, forming a carcinoecium around the crustacean's abdomen; the families of this group are Epizoanthidae, Gerardiidae, and Parazoanthidae.

Ecology. Like all cnidarians, zoanthids are carnivorous. Bits of food dropped by their hosts may be important to those living with hermit crabs. Zooxanthellate zoanthids may derive much of their energy from their photosymbionts. Zoanthids of at least one species are bioluminescent. One species of *Palythoa* forms a nonproteinaceous toxin that was ostensibly used by Hawai'ian warriors to tip their spears; unlike other cnidarian toxins, palytoxin is not associated with nematocysts, but seems to be produced within the gastrovascular cavity, perhaps by symbiotic microorganisms, and functions to protect the eggs from predation.

Reproduction. Some zoanthids are hermaphroditic (individuals are both male and female) and some gonochoric (individuals are either male or female); most broadcast their gametes but some brood. Unique to zoanthids are Semper's larvae, of which there are two types: the zoanthina, with a girdle of long cilia and 12 mesenteries, is characteristic of the genera *Isaurus* and *Zoanthus*; the zoanthella, with a ventral band of long cilia extending from the oral to the aboral end, is characteristic of the genera *Palythoa*, *Protopalythoa*, and *Sphenopus*. Other zoanthids produce pelagic larvae other than Semper's. Daughter polyps of a colony arise by budding from the stolon or the polyp base, or by longitudinal or fission; transverse fission occurs in the solitary zoanthid *Sphenopus*.

Phylogeny. The phylogenetic position of the Zoanthidea is obscure. These animals have some features reminiscent of the Paleozoic rugosan corals, but because nothing is known about the soft-tissue morphology of the latter, and the former do not form skeletons, which are all that is known of rugosans, it is impossible to determine if this resemblance is phylogenetically significant or just coincidence. Molecular data support the Antipatharia (black corals) as the nearest extant relative of zoanthids. *See* ANTHOZOA.

Daphne G. Fautin

Bibliography. H. Erhardt and D. Knop, Order Zoantharia, *in Corals: Indo-Pacific Field Guide*, pp. 269–278, Ikan, Frankfurt, 2005; J. P. Hoover, Zoanthids (colonial anemones), in *Hawai'i's Sea Creatures: A Guide to Hawai'i's Marine Invertebrates*, pp. 42–45, Mutual Publishing, Honolulu, 1998; R. L. Manuel, Order Zoantharia Gray, 1870, *in British Anthozoa*, pp. 70–83, Linnaean Society, London, 1870.

Zodiac

The band of sky through which the Sun, Moon, and planets apparently move in the course of the year. The Babylonians, about 2500 years ago, divided the zodiac into 12 parts, which correspond to constellations. These zodiacal constellations, in order around the sky, are Aries, the Ram; Taurus, the Bull; Gemini, the Twins; Cancer, the Crab; Leo, the Lion; Virgo, the Virgin; Libra, the Scales; Scorpius, the Scorpion; Sagittarius, the Archer, Capricornus, the Sea Goat; Aquarius, the Water Carrier; and Pisces, the Fish. These constellations are based on Greek myths.

Because of the precession of the equinoxes, the positions of the constellations in the sky have drifted from the dates of the year with which they were associated thousands of years ago. Thus the popular astrological "signs" of the zodiac are not actually those that currently correspond to the sky. Though the vernal equinox is often called the first point of Aries, precession has moved it into Pisces, not far from Aquarius. Contemporary astrologers claim not to base detailed horoscopes on these outmoded zodiacal dates; nonetheless there is no basis for astrology, and tests have shown that astrological predictions are no more valid than chance or than profiles based on psychology dealing with those receiving the horoscopes. *See* PRECESSION OF EQUINOXES.

The Sun actually passes through parts of 13 constellations, as currently defined. Also, if the zodiac is defined as the region within latitudes $\pm 8^{\circ}$, which accommodates the eight planets through Neptune,

it contains all or part of 24 constellations. *See* CONSTELLATION; ECLIPTIC. Jay M. Pasachoff

Bibliography. M. Lachièze and J.-P. Luminet, *Figures du Ciel*, Bibliothèque Nationale de France, Seuil, 1998; J. Meeus, *More Mathematical Morsels*, William-Bell, 2002; J. Mosley, The real, real constellations of the zodiac, *The Planetarian*, 28(4):5, December 1999; J. M. Pasachoff, *Field Guide to the Stars and Planets*, 4th ed., Houghton Mifflin, Boston, rev. 2003; C. Stott (ed.), *Images of the Universe*, Cambridge University Press, 1991.

Zodiacal light

A diffuse, night-sky luminosity easily seen at low to middle geographic latitudes in the absence of moonlight. It is caused by sunlight scattered and absorbed by interplanetary (solar system) dust particles. Zodiacal light extends over the entire sky, but it is brightest toward the Sun and in the zodiacal band. (In the ecliptic at 30° from the Sun its visual brightness is three times that of the brightest part of the Milky Way.) It is best seen in the west after evening twilight and in the east before morning twilight (**Fig. 1**), when the ecliptic is close to the vertical. In the Northern Hemisphere this corresponds to spring evenings and autumn mornings. *See* ECLIPTIC; INTERPLANETARY MATTER.

The visual zodiacal light brightness decreases monotonically with elongation (angular distance from the Sun) to a relatively flat minimum in the ecliptic at 120 to 140°, after which it gradually increases out to the antisolar point at 180°. The corresponding polarization reaches a maximum of approximately 20% at 70° elongation, falling to zero at 160°. The polarization changes direction between 160 and 180°, where it is again zero. Such characteristics of brightness and polarization, when combined with information on color and angular dependence in and out of the ecliptic can be used to provide information on the optical properties and spatial distribution of the dust. Ground and space observations have shown the zodiacal light to have approximately the same color as the Sun. It has been shown to be remarkably constant (to 1%) over periods as long as a solar cycle (11 years). Exceptions are small seasonal variations due to the motion of the observer with respect to the plane of the dust and occasional plasma radiation, mostly light associated with coronal mass ejections (CMEs). The density of particles responsible for the zodiacal light falls off somewhat faster than $1/R$, where R is heliocentric distance. The particles are primarily in the size range of tens to hundreds of micrometers in diameter, and are believed to originate primarily from comets. Band structure seen in infrared measurements of thermal emission from the dust may be due to particles arising from collisions in and near the asteroid belt. Ultraviolet and, especially, infrared observations from space are providing information that will lead to a better understanding of the overall physics of the interplanetary dust. *See* ASTEROID; COMET; INFRARED ASTRONOMY; SOLAR WIND; ULTRAVIOLET ASTRONOMY.

Fig. 1. Morning zodiacal light as it typically appears at a low, northern latitude site, Mount Haleakala, Hawaii, at an altitude of 10,000 ft (3300 m). The outcropping of rock at the base of the zodiacal cone is approximately 20° from the Sun. (*P. B. Hutchison*)

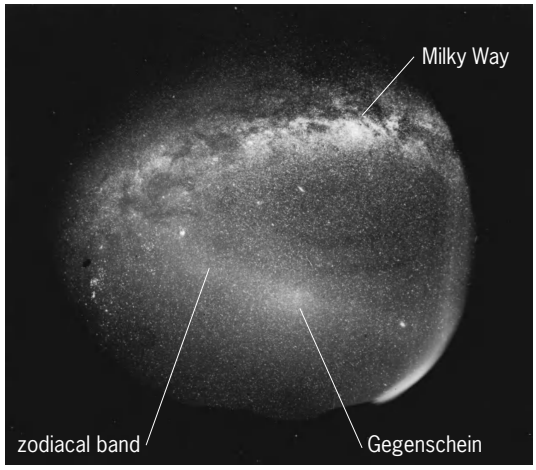


Fig. 2. All-sky photograph, taken from Mount Komagatake, Japan, showing the zodiacal band and Gegenschein, and the Milky Way.

Gegenschein. The enhanced brightness near the antisolar point is called the Gegenschein or counterglow (**Fig. 2**). It is barely above the visible threshold and is described by visual observers as being oval in appearance, 6 by 10° or larger, with the long axis in the ecliptic. Space observations have shown the Gegenschein to be an intrinsic part of the zodiacal light, rather than a separate phenomenon due to the Earth's atmosphere or to a concentration of dust opposite the Earth from the Sun; that is, the zodiacal dust particles have an increased scattering efficiency near backscattering.

Fraunhofer corona. The solar corona has two principal parts: a K-corona, which is due to Thomson scattering by free electrons located very close to the Sun, and an F-corona (also referred to as Fraunhofer or false corona or inner zodiacal light), which arises from scattering by zodiacal dust seen at small elongations. Although the Earth's atmosphere complicates attempts to observe zodiacal light closer than approximately 30° to the Sun (equivalent to seeing no particles closer than 0.5 astronomical unit to the

Sun), eclipse and space observations have shown the brightness to increase smoothly all the way in to the F-corona. The F-corona or inner zodiacal light and the "primary" zodiacal light seen at larger elongations come primarily from dust particles located relatively close to the Sun. Near-Sun multicolor (ultraviolet to visual to infrared) observations from space can be expected to add significantly to knowledge of the physics, dynamics, and chemistry of dust near the Sun and in the transition to interplanetary space. *See* SOLAR CORONA; SUN. J. L. Weinberg

Bibliography. E. Grün et al. (eds.), *Interplanetary Dust*, Springer, 2001; C. Leinert et al., The 1997 reference of diffuse night sky brightness, *Astron. Astrophys. Suppl. Ser.*, 127:1–99, 1997; Proceedings of Zodiacal Cloud Sciences, Kobe, Japan, September 1997, *Earth Planets Space*, 50:463–610, 1998.

Zone refining

One of a number of techniques used in the preparation of high-purity materials. The technique is capable of producing very low impurity levels, namely, parts per million or less in a wide range of materials, including metals, alloys, intermetallic compounds, semiconductors, and inorganic and organic chemical compounds. In principle, zone refining takes advantage of the fact that the solubility level of an impurity is different in the liquid and solid phases of the material being purified; it is therefore possible to segregate or redistribute an impurity within the material of interest. In practice, a narrow molten zone is moved slowly along the complete length of the specimen in order to effect the impurity segregation. *See* SOLUTION; SOLVENT.

Impurity atoms either raise or lower the melting point of the host material. There is also a difference in the concentration of the impurity in the liquid phase and in the solid phase when the liquid and solid exist together in equilibrium. In zone refining, advantage is taken of this difference, and the impurity atoms are gradually segregated to one end of the starting material. To do this, a molten zone is passed from one end of the impure material to the other, as in **Fig. 1***a*, and the process is repeated several times. The end to which the impurities are segregated depends on whether the impurity raises or lowers the melting point of the pure material; a lowering of the melting point is more common, in which case impurities are moved in the direction of travel of the molten zone.

The extent or effectiveness of the impurity distribution is controlled by the distribution coefficient, k, defined as the ratio of the impurity concentration in the solid to that in the liquid at a given temperature. The effect of multiple zone passes (in the same direction) on impurity content along the material is illustrated in Fig. 1*b*.

Zone leveling is a particular form of zone refining. Here zone passes are made in alternate directions so that a uniform (homogeneous) distribution of impurities (that is, alloying) is produced along the material. In this way desired levels of deliberately added impurities may be achieved in this form of zone refining.

In commercial application the object is to optimize the variables in order to produce the required amount of material having the necessary level of purity with a minimum of time and expense. Primary variables are zone length, interzone spacing in a multiple zone batch unit, number of zone passes, and travel rate of the zone along the material. The choice of zone length is governed by practical considerations and by the physical properties of the material; a narrow molten zone with a sharp demarcation between liquid and solid phases is desired. Small numbers of zone passes reduce the total time required but at the expense of the usable length of material. For large numbers of zone passes, the best segregation occurs for small zone lengths. Increasing the travel speed decreases the time per pass, but there is a concomitant decrease in the efficiency of segregation. Typical conditions for the zone refining of aluminum to a purity better than 99.9995% are a zone speed of 1 in./h (2.5 cm/h), 30 zone passes, and a zone length to material lenght ratio of 0.1.

The many designs of zone-refining apparatus center on factors such as the form of heating (or cooling), the form of container (material or shape), the travel mechanism, stirring of the liquid in the molten zone, and protective atmosphere. Common modes of heat input for material in a container are resistance heating and induction heating (3 Hz to 6 MHz in the 5- to 50-kW power output

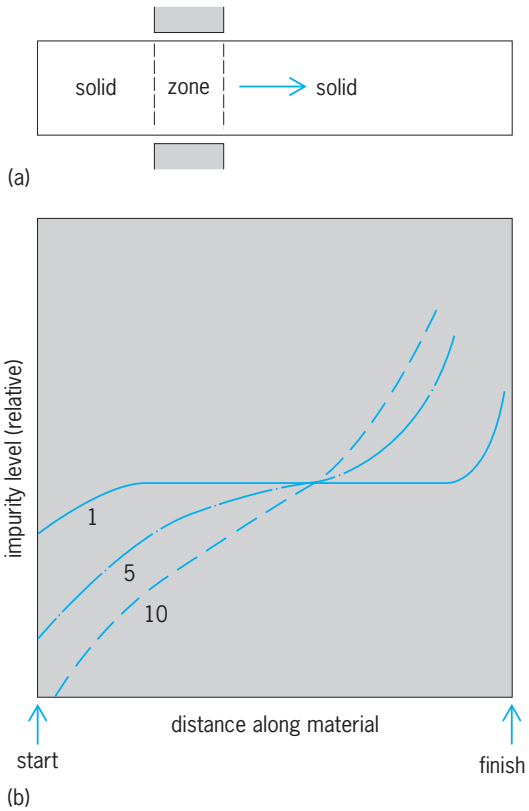


Fig. 1. Zone refining. (*a*) Passage of a molten zone along the material to be purified. (*b*) Effect of 1, 5, and 10 zone passes on the impurity distribution along the material.

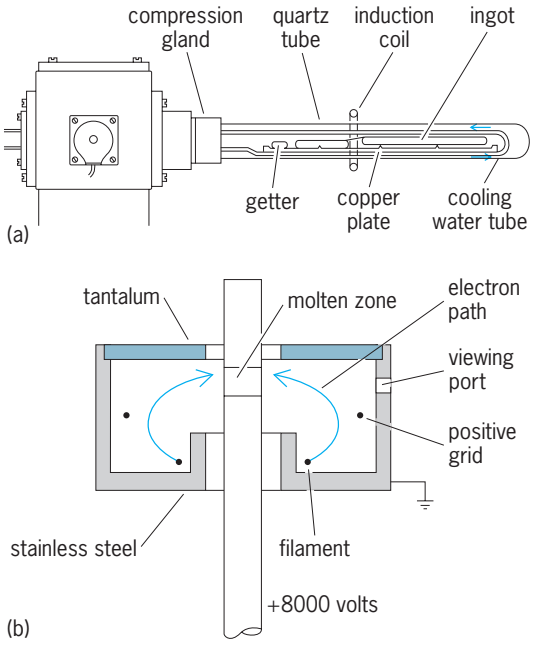


Fig. 2. Industrial techniques. (*a*) Zone refining on a water-cooled hearth. (*b*) Electron-beam zone-refining arrangement. (*After R. F. Bunshah, ed., Techniques of Metals Research, vol. 1, John Wiley and Sons, 1968*)

range). When the material to be purified melts below room temperature, it is necessary to zone-refine in a refrigerated atmosphere or to traverse a solid zone along the bar or column of liquid material. Container materials in use are plastics, metals, glass, graphite, and oxides of aluminum, zirconium, magnesium, silicon, and beryllium. The choice becomes more restrictive as the temperature of the molten zone increases. The upper limit of melting point for metals and alloys to be zone-refined in a container is ~2700°F (1500°C), since excessive reaction occurs with the container above this temperature. Low-melting-point organic compounds are usually purified in glass containers or plastics. High-melting-point or reactive metals, compounds, and semiconductors may be zone-refined in a container if the latter is water-cooled, as in **Fig. 2***a*. *See* RESISTANCE HEATING.

In floating zone refining, a molten zone is held in place by its own surface tension between two vertical colinear rods. Since no container is needed, contamination problems are avoided. The most common form of heating is electron bombardment (Fig. 2*b*), or induction heating. In this way refractory metals and compounds can be zone-refined. *See* INDUCTION HEATING.

The importance of the control and improvement of purity lies in the fact that, in most cases, the behavior of a material is a direct function of impurity content; the material behavior is manifested in mechanical, electrical, chemical, thermal, and optical properties. *See* METALLURGY. Alan Lawley

Bibliography. A. Choudhury, *Vacuum Metallurgy*, 1990; *Kirk-Othmer Encyclopedia of Chemical Technology*, 4th ed., 1997; W. G. Pfann and D. T. Hawkins, *Zone Melting*, 1966, reprint 1978.

Zooarcheology

Zooarcheology or archeozoology is the study of animal remains from archeological sites. Most such remains derive from people's meals. In other words, zooarcheology is essentially the study of ancient garbage, mainly bones and teeth of mammals, as well as birds, fish, mollusks, and even insects. Zooarcheology helps to provide a more complete picture of people's environment and way of life, especially their economy, as reflected in the relationship between people and animals 100,000 or even just 100 years ago.

Methods. The first task of the zooarcheologist is the taxonomic and anatomical identification of the faunal remains, ideally with the aid of a reference collection of modern identified specimens. Bone measurements help identify species size changes in the archeological record, and they can help to separate closely related species, such as sheep and goats, and domestic animals and their wild ancestors. Age-at-death in many mammals can be determined by the state of eruption and wear of the teeth and fusion of long-bone epiphyses.

A zooarcheological study is limited by the number of remains. Thus, 10 bones will reveal which species were exploited; 100 bones can reveal roughly in what proportion they were exploited; and 1000 bones may provide intraspecific information, such as the proportion of different age groups and sexes culled from the herd. In a sample of 10,000 bones, there should be enough mandibles to plot a continuous age distribution of the herbivores.

Bones are usually damaged from butchery, trampling, cooking, and gnawing in antiquity. Many have been subject to varying degrees of erosion in the soil. (These environmental alterations are included in the subject of taphonomy.) An archeological assemblage of bones, therefore, comprises numerous broken limb bone ends, isolated teeth, and unidentifiable fragments. The method of excavation can also have an effect, especially when the soil is not sieved, leading to the poor recovery of smaller bones and isolated teeth.

Environmental reconstruction. Animal remains often indicate global climate patterns of the past. For instance, in cold periods (such as the Ice Ages), cold-weather species were distributed in what are temperate areas today. Given sufficiently large samples in a multiperiod site, varying frequencies of species can reflect climate changes. At Combe-Grenal, France, Francois Bordes and Prat interpreted geologic levels with large numbers of reindeer as having formed during cold periods, layers with more red deer and wild cattle as having formed during warmer periods. At Boomplaas cave in South Africa, Richard Klein studied the animal remains and found that the lowermost and uppermost levels date back to the last Interglacial and present Interglacial (Holocene), respectively. These levels contain a predominantly woodland and bush fauna with hyrax and baboons. The intermediate levels date between 20,000 B.C. and 10,000 B.C. and were

found to contain a fauna consisting predominantly of zebras and wildebeest. These are more typical of grassland and, according to Klein, reflect cool, dry conditions which prevailed in South Africa during the late glacial period. An increase or decrease in species diversity may reflect milder or harsher environmental conditions, respectively. *See* CAVE.

The size of mammal bones and teeth can indicate temperature change. Today, mammals are often larger in colder areas (Bergmann's rule). In several zooarcheological assemblages (including South Africa and Israel) spanning the end of the last Ice Age and Holocene period (12,000 years ago), many mammals became smaller as the temperature increased. *See* CLIMATE HISTORY.

Hunting. It is clear that toward the end of the Pleistocene, people were effective hunters of big game. They occupied most of the continental landmass and subsequently oceanic islands as well. The arrival of hunting people in Australia, the Americas, and many islands coincided with the extinction of many large animals such as the giant kangaroo, giant wombat, and diprotodon in Australia; the ground sloth, giant beaver, horse camel, mammoth, and mastodon in North America; the moas in New Zealand; and the pygmy elephant and hippotamus on some Mediterranean islands. This catastrophic loss of fauna has been called "megafaunal overkill."

Zooarcheology seeks evidence of an association between early humans invaders and the extinction of the animals. In the United States, there are a number of localities where, for example, mammoth remains are associated with human artifacts dated between 9200 and 9000 B.C. However, such evidence is scarce, presumably because extinctions happened rapidly and the coexistence of people and megafauna was brief. Although it is debated whether climate change or overhunting (or both) was responsible for the extinctions in America, overhunting was probably the main cause.

In Africa and Eurasia, where humans and animals had coevolved for a long time, extinction was less abrupt. However, by 9000 years ago in the Near East, zooarcheological studies indicate that people exploited fewer large animals—which had presumably become scarce or even extinct. People exploited more small mammals such as hares, fish, and birds. The mortality curves of gazelles hunted in the Levant between 50,000 and 10,000 years ago show that by the end of this period, greater numbers of young were being hunted. This and the shift to hunting of smaller animals may reflect increased pressure on the resources by increased numbers of people. Continued demographic pressure subsequently forced people to husband animals (and plants) and colonize new territories such as oceanic islands. *See* PLEISTOCENE.

Domestication. One challenge for zooarcheologists is to try and discover when animals were first domesticated and why. The shift from hunting to husbanding may be recognized zooarcheologically in four ways:

1. *Size change.* Species size changes can occur due to human selective breeding practices. For example, contempoary dogs, cattle, sheep, and pigs are smaller than their wild ancestors: wolves, aurochs (wild cattle), mouflon sheep, and wild boar. However, domestic guinea pigs are larger than wild guinea pigs. Care is needed to avoid confusing size variation due to environmental change with size changes caused by human intervention. Size can also vary with sex and age, so these variables need to be controlled when considering measurements to distinguish between bones of wild and domestic mammals.

2. *Importation of exotic species.* Before the sixth millennium B.C., there were no sheep in Europe. Sheep bones found on sixth-millennium archeological sites in Europe must, therefore, derive from introduced and thus, domesticated animals. A similar example is the shipment of sheep, goats, cattle, cats, and dogs to Cyprus in the eighth millennium B.C.

3. *Species spectrum change.* In several parts of the Near East, a shift from gazelle to sheep and goats occurred in the seventh millennium B.C. At Franchthi cave in Greece, Sebastian Payne noticed a shift from horse, deer, wild boar, and ibex to sheep (previously unknown in Greece), goats, pigs, and cattle in the sixth millenium B.C. Both Near Eastern and Greek faunal shifts probably signify the change from hunting to husbanding.

4. *Cultural signs.* Several burials of dogs with humans have been excavated in Israeli Natufian (an Epipaleolithic culture) sites. They date back to the tenth millennium B.C., and they are the earliest known burials of this kind. At a Mesolithic site in Switzerland, Louis Chaix interpreted a pathological deformation between the first and second molar teeth of a brown bear as caused by a leather thong, indicating that it was probably kept as a pet. Shed caprine (goat) milk teeth (first teeth) in early Neolithic Greek and French sites suggest penning of sheep and goats.

The earliest evidence for domesticated animals in the Old World is of dogs in the Near East about 10,000 B.C., followed one or two thousand years later by the food animals such as sheep, goat, cattle, and pig. The origins of domesticated animals in eastern Asia (such as the pig and chicken) and the New World (including the llama, alpaca, and guinea pig) are less well known. *See* ARTIODACTYLA; DOGS; NEOLITHIC; PALEONTOLOGY.

Secondary uses of animals. Patterns of exploitation of animals for uses other than meat—for example, sheep and goats for their wool and milk—is more difficult to identify since wool and milk are rarely preserved. Thus a more indirect approach is adopted. Since in a meat-producing economy animals are slaughtered in their prime, high numbers of juvenile bones and teeth are found in an archeological assemblage. If wool is the prime objective, slaughter may be delayed until animals are 5 or more years of age, when teeth are well worn and epiphyses fused. There is some evidence, again from the Near East, of a shift in sheep and goat culling patterns in the fifth or fourth millennium B.C., perhaps reflecting the beginning of wool or milk exploitation.

Power is another secondary use of animals. Equids, camelids, and cattle have long been an important

source of power for transport and plowing. Excessive strain in life can cause joint disease, sometimes recognized as bony outgrowths and fusions of foot bones, though old age or soft soil may also induce the same arthropathies. Properly controlled samples of modern animals used for traction are needed to test this assumption.

Other uses of animals. Animal remains can sometimes provide other kinds of useful information about the human inhabitants of a site. Remains of migratory birds (for examples, winter visitors) can provide information such as the season an animal resource was exploited. Animals or specific parts of them were often sacrificed and used ritually. For example, the body of a man buried in Bronze Age Northamptonshire, England (Irthlingborough), was covered with 185 cattle skulls and the mandibles, scapulae, and pelves of 20 or 30 cattle. No other parts of the animal skeletons were present. Bones were sometimes used as construction materials. For example, in the Ice Age of Ukraine where wood was scarce, mammoth bones and tusks were used to construct the scaffolding for shelters, and presumably then covered with skins.

Overview. Zooarcheology is a multidisciplinary endeavor requiring knowledge of anatomy and biometry as well as an appreciation of the archeological questions that need to be addressed. Unlike most paleontological collections, zooarcheological collections are usualy well dated and comprise large numbers of bones. They provide excellent opportunities to study microevolution. However, much remains to be explored in this relatively new science. *See* ANTHROPOLOGY; ARCHEOLOGY; FOSSIL HUMANS; ZOOLOGY.

Simon J. M. Davis

Bibliography. S. J. M. Davis, *The Archaeology of Animals*, Yale University Press, New Haven, 1987; R. G. Klein and K. Cruz-Uribe, *The Analysis of Animal Bones from Archaeological Sites*, University of Chicago Press, 1984; P. S. Martin and R. G. Klein (eds.), *Quaternary Extinctions: A Prehistoric Revolution*, University of Arizona Press, 1984; E. J. Reitz and E. S. Wing, *Zooarchaeology*, Cambridge University Press, 1999; B. Wilson, C. Grigson, and S. Payne (eds.), *Ageing and Sexing Animal Bones from Archaeological Sites*, BAR British Series, 109, Oxford, 1982.

Zoogeography

The subdivision of the science of biogeography that is concerned with the detailed description of the distribution of animals and how their past distribution has produced present-day patterns. Scientists in this field attempt to formulate theories that explain the present distributions as elucidated by geography, physiography, climate, ecological correlates (especially vegetation), geological history, the canons of evolutionary theory, and an understanding of the evolutionary relationships of the particular animals under study. Zoogeographical theories are then tested by new data from all germane fields to amplify, verify, or falsify the constructs. In this sense, zoogeography is an integrative science that synthesizes data from other disciplines to apply to the realities of animal distribution.

Realities. The field of zoogeography is based upon five observations and two conclusions. The observations are as follows. (1) Each species and higher group of animals has a discrete nonrandom distribution in space and time (for example, the gorilla occurs only in two forest areas in Africa). (2) Different geographical regions have an assemblage of distinctive animals that coexist (for example, the fauna of Africa south of the Sahara with its monkeys, pigs, and antelopes is totally different from the fauna of Australia with its platypuses, kangaroos, and wombats). (3) These differences (and similarities) cannot be explained by the amount of distance between the regions or by the area of the region alone [for example, the fauna of Europe and eastern Asia is strikingly similar although separated by 6900 mi (11,500 km) of land, while the faunas of Borneo and New Guinea are extremely different although separated by a tenth of that distance across land and water]. (4) Faunas strikingly different from those found today previously occurred in all geographical regions (for example, dinosaurs existed over much of the world in the Cretaceous). (5) Faunas resembling those found today or their antecedents previously occurred, sometimes at sites far distant from their current range (for example, the subtropical-warm temperate fauna of Eocene Wyoming, including many fresh-water fishes, salamander, and turtle groups, is now restricted to the southeastern United States).

The conclusions are as follows. (1) There are recognizible recurrent patterns of animal distribution. (2) These patterns represent faunas composed of species and higher groups that have evolved through time in association with one another.

Approaches. Two rather different approaches have dominated the study of zoogeography since the beginning of the nineteenth century: ecological and historical. Ecological zoogeography attempts to explain current distribution patterns principally in terms of the ecological requirements of animals, with particular emphasis on environmental parameters, physiological tolerances, ecological roles, and adaptations. The space and time scales in this approach are narrow, and emphasis is upon the statics and dynamics of current or very recent events.

Historical zoogeography recognizes that each major geographical area has a different assemblage of species, that certain systematic groups of organisms tend to cluster geographically, and that the interaction of geography, climate, and evolutionary processes over a long time span is responsible for the patterns or general tracks. Emphasis in this approach is upon the statics and dynamics of major geographical and geological events ranging across vast areas and substantial time intervals of up to millions of years. The approach is based on concordant evolutionary association of diverse groups through time.

Ecological zoogeography is the study of animal distributions in terms of their environments; historical zoogeography is the study of animal distributions in terms of evolutionary history.

Comparison of three grassland biomes

Ecological roles	North American	African	Australian
Top predator	Wolf (dog)	Lion (cat)	Tasmanian wolf (marsupial)
Large herbivore	Bison (cattle)	Zebra (horse)	Red kangaroo (marsupial)
Small predator-scavenger	Coyote (dog)	Hyena (hyena)	Tasmanian devil (marsupial)

Ecological zoogeography. Ecological zoogeography is rooted within the discoveries of nineteenth-century plant ecologists and physiologists. They found that under similar conditions of temperature and moisture, terrestrial plants develop similar growth forms, regardless of their evolutionary relationships, to produce one of the following vegetation forms: forest, woodland, savanna, grassland, or scrub. Subsequently it was demonstrated that these units showed differentiation associated with major soil types and broad latitudinal climatic regions. Each of these main kinds of vegetation is called a formation type (such as tropical lowland evergreen forest), and its geographical subdivisions are formations (such as South American tropical lowland evergreen forest; African tropical lowland evergreen forest).

Later investigators realized that within each formation a series of animals had evolved to undertake homologous ecological functions in the dynamics of the community, so that the concept of biome-type (vegetation and associated animals) was developed. The biome-type is a series of major geographical climatic regions characterized by similar ecological adaptations in plants and animals. For example, the grassland biome-type may be typified by the comparisons in the **table**. The biome-types are distinctive from the zoological point of view in that ecological equivalents (such as top predator) in the different biomes are usually from phylogenetically nonrelated stocks. M. D. F. Udvardy provides the best recent summary of the general distribution of world terrestrial biomes.

The concept of biomes is a method of generalizing the distribution of animals by major environments on a latitudinal basis. The ecological zoogeographer is also interested in vertical (altitudinal and bathymetric) zonation as a feature of animal distribution. The idea of zonation is based upon the recognition that there are ecoclinal gradients in the parameters of the environment (especially temperature) with an increase in altitude or in ocean depth. The composition of species distributions along these ecoclines, because of differences in ecological requirements, produces recognizable and characteristic life zones. As an example, **Fig. 1** shows the principal life zones in the ocean. These zones are divided into two groups: benthos or substrate zones and pelagic or free-swimming zones. Altitudinal life zones on land have been similarly described. *See* BIOME; LIFE ZONES.

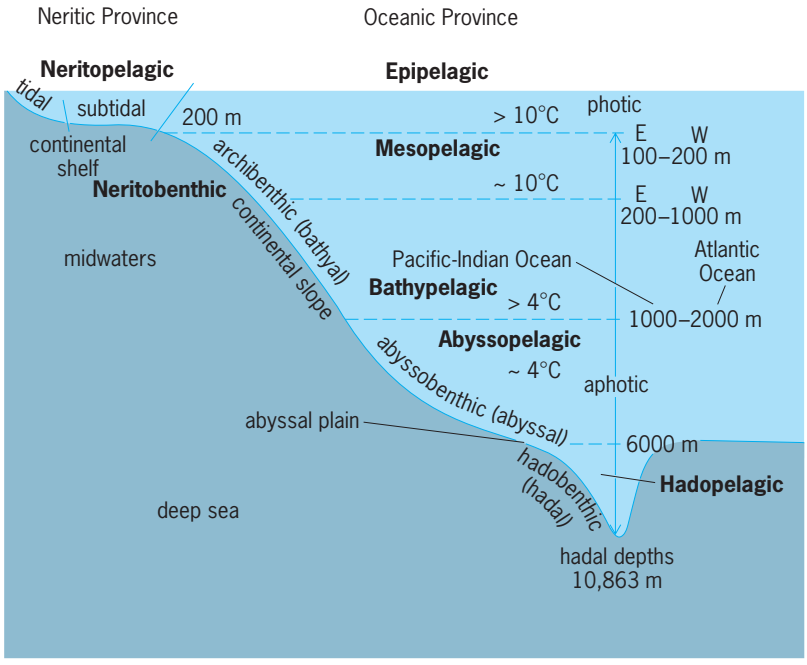


Fig. 1. Life zones in the sea. 1 m = 3.3 ft; °F = (°C × 1.8) + 32.

Historical zoogeography. This approach has its origins in systematic biology. Workers in this field recognized very early in the nineteenth century that different geographical regions support different faunas and that representatives of these faunas occur in a wide variety of environments (for example, apes and monkeys in Africa in rainforest to desert environments). In addition, they noted that different systematic groups tend to cluster geographically (kangaroos in Australia and anteaters in South America). With the development of the canons of evolutionary theory later in the century, a framework for understanding the evolution of faunas through time became the backbone of historical zoogeography.

The raw data of historical zoogeography are the distributions or tracks of individual species of animals in space (geographical ecology) and time. Because each species has its own set of peculiar ecological requirements and its own unique evolutionary history, each has a discrete nonrandom ecogeographical distribution. As a consequence, no species is universally present, and many species have small or unique tracks. The first level of generalization in zoogeography is based on the recognition that, in spite of the unique nature of individual species distributions, many tracks are concordant or show a common pattern. Determination of patterns involving the coincident distribution of species or monophyletic groups (genera, families, and so on) of species (generalized tracks) is the fundamental step in zoogeographical analysis.

The second level of generalization in this process

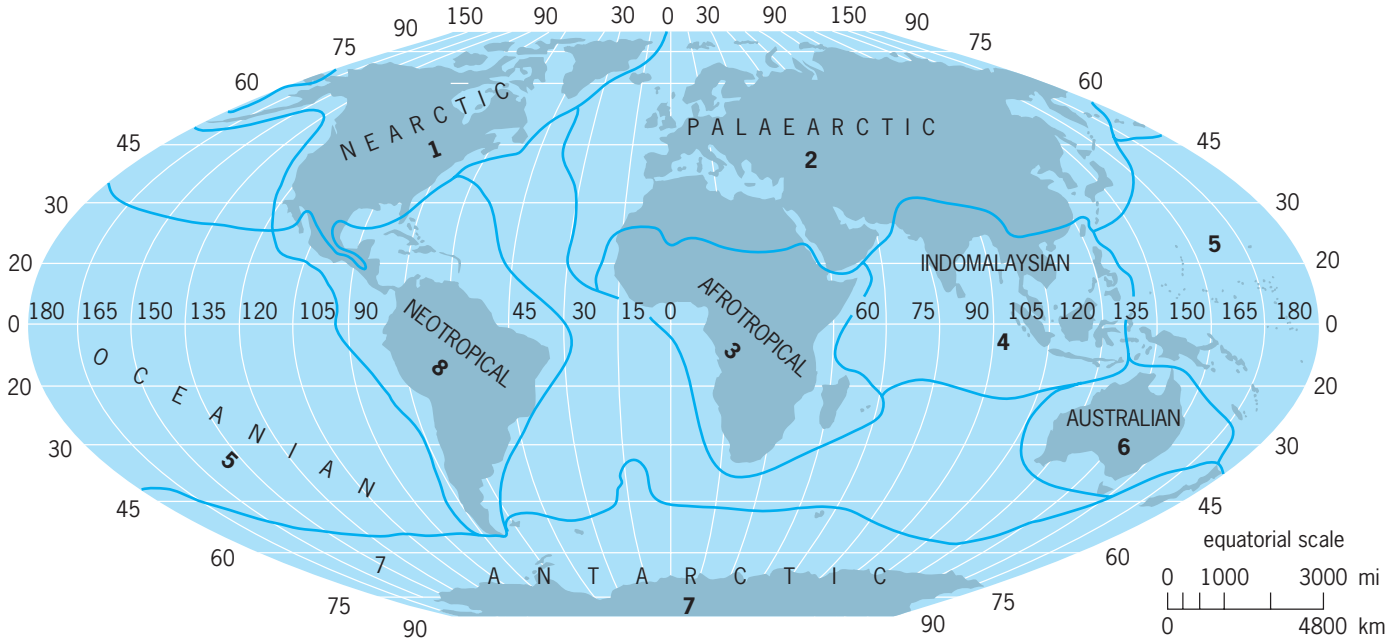


Fig. 2. Terrestrial biogeographic realms of the world. (*After M. D. F. Udvardy, A classification of the biogeographical provinces of the world, Int. Union Conserv. Nat. Occas. Pap., 18:1–49, 1975*)

is to cluster the strongly recurrent generalized tracks involving extensive geographical areas, whose components are then regarded as the major modern faunas. A third level of generalization attempts to tentatively identify the historical source units (ancestral faunas) that have contributed to the modern patterns.

Zoogeographical patterns. The early workers in zoogeography, especially P. Sclater in 1858 (birds) and A. Wallace in 1876, developed a system for classifying major patterns of terrestrial distributions, which in modified form (**Fig. 2**) is still applicable today. This system reflects the long-term isolation of major parts of the Earth's surface and the consequent divergent evolution of the fauna in each isolate. In some cases, subsequent reconnections of formerly isolated areas have blurred the distinctiveness of the core faunas in an area of transition. The major units in this system are called biogeographic realms; a series of biogeographic provinces form subdivisions within the larger divisions.

The eight recognized realms (which correspond to what were called zoogeographic regions in the Sclater-Wallace system and by many subsequent authors), with the number of provinces recognized for each realm indicated in parentheses are briefly characterized below. These units are based primarily on comparisons for the best-known animal groups. For this reason the several classes of vertebrates are used to characterize the realms below, although terrestrial and fresh-water invertebrates follow similar patterns but are less well known.

Nearctic Realm (22). North America, north of the edge of the Mexican-Guatemala highlands, although many workers would place the boundary with the Neotropical Realm at the edge of the Mexican plateau, west of the Isthmus of Tehuantepec. In terms of vegetation the northern areas support transcontinental zones of tundra and needle-leaf coniferous forests (taiga). The eastern portion of the continent to the south of the taiga zone was originally covered with broadleaf deciduous forests, and the central region by extensive grasslands. Coniferous forests predominate along the northwest coast and in the Rocky Mountains, Sierra Nevada, and upper regions of the Sierra Madre of Mexico. Much of the southwestern United States and adjacent Mexico are covered by oak savannas, scrub, and desertic vegetation. In terms of the fauna, a great many groups and species are shared with the Palearctic (that is, they have a holarctic distribution), and in more southern areas with the Neotropical Realm.

Distinctive vertebrate groups of the Nearctic include bowfin, gar-pike, bullhead catfishes, panfish, and basses; hellbenders, mud puppies, amphiumas and sirens, the bell-frog (*Ascaphus*), spadefoot toads, snapping turtles, the land tortoises (*Gopherus*), the Gila monster and its ally the Mexican beaded lizard (*Heloderma*), the night lizards, fence lizards and their allies, glass lizards, worm lizard (*Rhineura*), rattlesnakes, and alligator; flycatchers, vireos, orioles, hummingbirds (all shared with tropical America); racoons, pocket gophers (also tropical), and pronghorn. Many characteristic birds and mammals (woodpeckers, wood warblers, tanagers, pikas, flying squirrels, beaver, jaguar, and cougar) belong to wide-ranging groups, with endemic species in the New World.

Palearctic Realm (44). The Eurasian landmass from southern China and the southern slopes of the Himalaya Mountains northward and westward from the Indus Valley to include the Arabian Peninsula and northern Africa. As with the Nearctic Realm, transcontinental zones of tundra and taiga occur in

the northernmost areas. Broadleaf deciduous forest originally occurred over much of western (Europe) and eastern (China) portions of the realm, and the central areas of the landmass were steppe, grassland, or desert. Scrub-to-desert vegetation is typical of the Mediterranean area east to the Indus. The Palearctic fauna is a peculiar mixture of widespread groups with strong affinities to the Nearctic and others from the Old World tropics. For this reason, there is very little endemism at the major group level, and the distinctive groups are ones with a major radiation of species in the Palearctic. Distinctive vertebrate groups include a vast radiation of cyprinid fishes and numerous salmonids; many salamanders, lacertid lizards, slowworms, vipers; warblers, hedge sparrows (endemic); dormice, jerboas, the two-humped camel, and pandas.

Africotropical Realm (29). Africa south of the Sahara Desert and Madagascar. The realm is mostly tropical with broadleaf evergreen forests along the margins of the Gulf of Guinea and inland through the Congo Basin and on the eastern coast of Madagascar. Deserts cover the Kalahari, Namib, and the Somali Coast. Much of the rest of the continent was originally covered by tropical broadleaf deciduous forests which degraded to scrub-forest or savanna-grassland under human use. Montane evergreen forest occurs on the mountains of east Africa.

Although the realm shares many groups with tropical Asia and the Palearctic, it has one of the most easily recognizable faunas. Distinctive vertebrate groups for the continent include bichirs (lungfishes); caecilians, clawed frogs (*Xenopus*), reed frogs; side-neck turtles, chameleons, monitor lizards, sungazers, pythons, cobras, vipers, crocodiles; ostrich, secretary bird, hammerhead, touracos, mousebirds, helmet shrikes; otter shrew, golden mole, elephant shrews, monkeys, chimpanzees, gorilla, hyenas, aardvark, hyraxes, hippopotamuses, and spring haas. There are many genera and species of big game: the cats, rhinoceroses, buffalo, antelope, and pigs are African endemics. Madagascar lacks many important Continental groups but has a unique radiation of microhylid and reed frogs, specialized iguanids and boas, the endemic elephant birds (now extinct), flightless mesoenatids, endemic rollers, philapittas, and vangas; tenrecs, and lemurs, unique endemic mammals.

Indomalasian Realm (27). The Asian landmass from the west margin of the Indus Valley and south of the Himalaya Mountains and southern China and adjacent islands including Taiwan, the Philippines, and the Lesser Sundas. Most of the realm originally supported tropical forests with evergreen forests in western peninsular India, much of Burma, southeast Asia, southern China, the Malay Peninsula, and the Indo-Malayan Archipelago. Deciduous vegetation covered most of India and some areas of southeast Asia. Most of the latter has been converted to human use and degraded into scrub, grassland, or desert.

Distinctive groups of the Indo-Malayan include a high diversity of cyprinid fishes, loaches, labyrinth fishes; caecilians, rhacophorid tree frogs, microhylids; many endemic emydid turtles; gavial; pythons, pipe-snakes, cobras, sea snakes, pit vipers. pheasants, fairy bluebirds and leaf birds; hairy hedgehog, flying lemur, and spiny dormice. Endemics include monkeys, gibbons, orangutan, tarsiers, an elephant, a tapir, two rhinoceroses, and many antelopes.

Oceanian Realm (7). The Pacific islands of Melanesia, Micronesia, Polynesia, Hawaii, New Caledonia, and New Guinea. These islands are covered with tropical vegetation and show strong affinities with the Indo-Malayan unit. Many workers include New Guinea in the Australian Realm because of its great similarity in vertebrates to the latter. The remaining islands have a high degree of endemism, very few distinctive major groups, and a depauperate vertebrate fauna. There are no truly fresh-water fishes, frogs occur only on the islands near New Guinea and on the Fijis, snakes occur naturally only as far east as Fiji, while geckos and skinks range over the entire area and native flightless mammals occur only out to the Solomons.

Australian Realm (13). For these purposes, Australia and Tasmania, although many zoogeographers would include New Guinea and adjacent islands to the Solomons. Most of Australia is covered by grasslands, scrub, and desert, with temperate forests along the eastern and southeastern lowlands (including Tasmania) and tropical forests in the northwest. The Australian Realm has a unique fauna including these distinctive groups: an endemic lungfish, an osteoglossid (other fresh-water fishes are all marine-derived), myobatrachid frogs, pelodryadid tree frogs, side-neck turtles, a vast radiation of geckos and agamid lizards, pygopodids, pythons, many elapid snakes; cassowaries, emus, megapodes, owlet frogmouths, lyrebirds, scrubbirds, flowerpeckers, honey-eaters, bell magpies, magpie larks, bower birds, and birds of paradise; spiny anteater, platypus, marsupials—Tasmanian wolf, Tasmanian devil, anteater, bandicoots, Australian opossums, koala, wombats, kangaroos, wallabies, and wallaroos.

Antarctic Realm (4). The Antarctic continent, New Zealand, and the subantarctic islands. New Zealand has extensive temperate forests of mixed evergreen and deciduous trees over much of the North and South islands, with grassland predominating on parts of South Island. The subantarctic islands are relatively barren with vegetation primarily of grasses, sedges, and annual composites. The Antarctic continent is covered by ice and snow. Only New Zealand is of interest here because of its depauperate but unique fauna. It has no truly fresh-water fishes. A single amphibian, the frog *Leiopelma*, and the tuatara (*Sphenodon*), the only survivor of a Cretaceous order of reptiles, occur there. There are no turtles or snakes (except marine ones), but several endemic genera and species of geckos and skinks. Unique birds include the extinct moas, kiwis, an extinct flightless goose, flightless rails, a flightless parrot, and a predatory parrot (the kea). Native land mammals are two bats.

Neotropical Realm (47). Lowland Mexico, Central America, South America, and the Antilles. Evergreen

forests predominate along the Atlantic slope of Mexico, Central America, Panama, northwestern South America, and the Amazon Basin. Extensive deciduous forests formerly covered the lowlands of western Mexico and Central America, the Greater Antilles, and much of southern Brazil. Grassland and scrub vegetation cover extensive areas in central South America south of the Amazon Basin, while desert conditions occur along the west coast of Chile-Peru. The high mountains of Central America, the Andes, and the planaltino of Peru-Bolivia support a wide variety of montane vegetation. Temperate rainforest occurs in southern Chile, while steppe conditions prevail over the Patagonian region of southern Argentina and extreme southern Chile.

It is no wonder that, with this diversity of habitats, the Neotropical Realm has the richest fauna of any region. The fresh-water fish fauna is extremely diverse, but represented by rather primitive stocks, characins, gymnotid eels, several endemic catfish families as well as endemic osteoglossids, a lungfish, and synbranchid eels; distinctive are caecilians, Surinam toads, and leptodacylid and microhylid frogs; reptiles are side-neck turtles, caimans, many iguanid and microterid lizards, boas, coral snakes, and pit vipers; birds include rheas, tinamous, New World vultures, screamers, cracids, the hoatzin, seriamas, trumpeters, the limpkin, the sun bittern, seed snipes, potoos, motmots, hummingbirds, puffbirds, jacamars, toucans, manakins, and cotingas; distinctive mammals are phalangers, armadillos, anteaters, sloths, tapirs, peccaries, porcupines, capybara, paca, agouti, marmosets, and New World monkeys.

A similar biogeographical scheme has been established for marine areas; J. C. Briggs has provided an up-to-date summary of the system.

Zoogeographical dynamics. The early workers in zoogeography concentrated their effort on the description of patterns of animal distribution and the clustering of patterns into large units of distribution by major environment or major fauna. These static approaches, while useful in phrasing and structuring data, have proved inadequate for understanding scientific zoogeography since they address only the first of four key elements: (1) recognizing common patterns of distribution; (2) analyzing these patterns to determine common ecological or evolutionary processes that produced the patterns; (3) using the patterns and processes as a prediction of patterns for as yet unstudied groups and (4) for as yet undiscovered geographical and evolutionary events.

The static approach ultimately had a stultifying effect on the development of zoogeography, since by the middle of the twentieth century descriptive and narrative zoogeography had run out of new ideas. Fortunately since the 1960s the science has been revitalized by a resurgence of interest in the dynamics of animal distribution. This interest has led to the development of two major new theories of zoogeographical explanation, one ecological (the dynamic equilibrium theory) and one historical (the vicariance theory). The development of the latter theory has forced a reexamination of the previously dominant historical zoogeographical explanation (the dispersion theory), most effectively expounded earlier in the century by W. D. Matthew and G. G. Simpson.

Dynamic equilibrium theory. This theory of zoogeography was developed during the 1960s primarily by R. H. MacArthur and E. O. Wilson. It originally was aimed at developing predictive mathematical models that would explain the differences in numbers of species on islands of differing sizes and differing distances from the closest mainland source areas. Since almost all habitats on the mainland are patchy in distribution as well (meadows, lakes, mountaintops, and so forth), the theory can also be applied to any discontinuous (insular) segments of the same environment type.

The central axiom of the theory is that the fauna of any disjunct ecological area is a dynamic equilibrium between immigration of new species into the area and extinction of species already present. Species number is thus constant over ecological time, while evolution will act gradually over geological time to increase the equilibrium number of species. From this base it is possible to construct equilibrium models that predict the interactions of distance from the source area, areal extent of the disjunct area, and immigration versus extinction (**Fig. 3**). Data from well-known insular faunas supported the value of these models. Subsequently, controlled defaunization experiments on small islands and study of the immigration process through time confirmed the predictive power of the theory.

Several conclusions may be derived from these models and have been confirmed in the field: (1) Distant disjunct areas will have fewer species than those close to a source area. (2) Small disjunct areas will have fewer species than large ones. (3) Distant disjunct areas take longer to reach equilibria when originally sterile or defaunated than do those close to the source area. (4) The smaller the area and the closer the area source to the area, the higher the turnover rate (change in species composition). The essential insight of equilibrium theory is that local extinctions and immigrations are relatively frequent events. Development of more complex equilibrium models and extension of the approach to other areas of zoogeography are actively under way and promise to produce exciting new views of distribution events.

Dynamic historical biogeography. This field is currently undergoing a major revolution of thought stimulated by new knowledge of global tectonics based on the now generally accepted theory of continental drift as outlined by R. S. Dietz and J. C. Holden. Simply stated, theorists now believe that the major continents were formerly welded together as a supercontinent (Pangaea) that began to rift apart in the Triassic, about 190 million years before present. By the Early Jurassic, northern (Laurasia) and southern (Gondwanaland) landmasses had drifted apart. During the Cretaceous these masses fragmented further, and in the Cenozoic several southern segments became attached to the northern continents (Africa to Eurasia, India to Eurasia, and South America to North America). *See* PLATE TECTONICS.

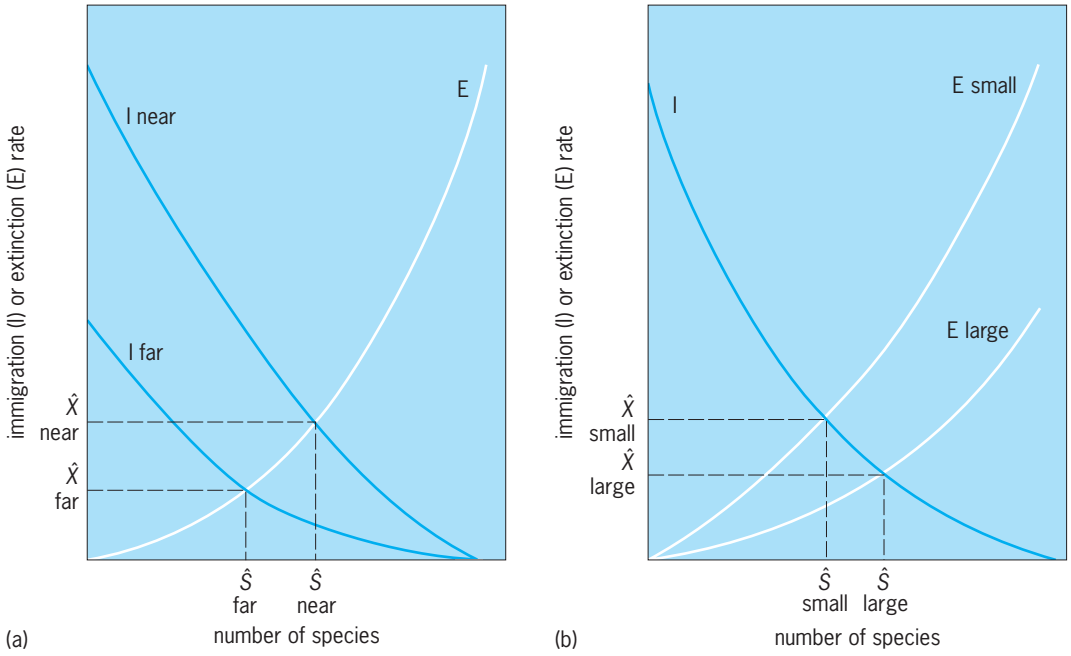


Fig. 3. A disjunct fauna is an equilibrium in ecological time between immigration of new species and extinction of those already present. (*a*) Distance effect; a near island has large equilibrium number of species ($\hat{S}$) and turnover rate ($\hat{X}$). (*b*) Area effect; a large island has larger $\hat{S}$ and smaller $\hat{X}$. (*After D. S. Simberloff, Equilibrium theory of island biogeography and ecology, Annu. Rev. Ecol. Syst., 5:161–182, 1974*)

Previously, historical zoogeography had been dominated by the ideas of Matthew and Simpson, who believed in the permanency of the ocean basins and continents. These authors developed the idea that major groups originated on the northern continents and dispersed southward. P. J. Darlington developed a slightly different point of view and suggested that major groups arose in the Asian tropics and dispersed elsewhere on the continental masses across land bridges or marine barriers by island hopping. According to these kinds of ideas, the present-day distribution of lungfishes (tropical Africa, South America, and Australia) involved origin in the Old World (Asia?) and immigration across land bridges or by island hopping to the southern landmasses.

A group of zoogeographers, L. Croziat, G. Nelson, and D. E. Rosen, much influenced by the role of continental drift in geography, have developed and extended some ideas originally put forward by Croziat into a new theoretical construct. This theory focuses not on the dispersion of organisms from centers of origin, but on the idea that current zoogeographical patterns are the result of the fragmentation of previously continuous tracks by major physiographical change. Thus the present distribution of lungfishes is the result of the fragmentation of a previously continuous track by the breakup of Gondwanaland in the Cretaceous Period.

Essentially, this last development creates a controversial dichotomy in zoogeographical thought. One view emphasizes the active movement (dispersal) of animals as the principal agent responsible for patterns; the other regards dispersion as unimportant and emphasizes the movement and fragmentation of the landmasses and the relative immobility of animals as responsible for patterns. The latter position is called the vicariance theory, in distinction to the dispersion theory. An active, ongoing, vigorous interchange of ideas, with serious reexamination and critique of both schools, is the present theme of zoogeography.

The essential features of the dispersion and vicariance theories are as follows.

Dispersion theory: (1) A monophyletic group arises at a center of origin. (2) Each group disperses from this center. (3) Substantial numbers of monophyletic groups followed the same dispersal route at about the same time to contribute to the composition of a modern fauna. (4) A generalized track corresponds to a dispersion route. (5) Each modern fauna represents an assemblage derived from one to several historical source units. (6) Direction of dispersal may be deduced from tracks, evolutionary relations, and past geodynamic and climatic history. (7) Climate or physiographical change provides the major impetus or opportunity for dispersal. (8) Faunas were shaped by dispersion across barriers and subsequent evolution in isolation. (9) Dispersion is the key to explaining modern patterns: related groups separated by barriers have dispersed across them.

Vicariance theory: (1) Vicariants (allopatric species) arise after barriers separate parts of a formerly continuous population. (2) Substantial numbers of monophyletic groups are simultaneously affected by the same vicariating events (geographical barrier formation). (3) A generalized track estimates the faunal composition and geographical distribution of an ancestral biota before it subdivided (vicariated) into descendant faunas. (4) Each generalized track represents a historical source unit. (5) Sympatry of generalized tracks reflects geographical overlap of

different faunas due to dispersal. (6) The primary vicariating events are changes in world geography (geodynamics) that subdivided ancestral faunas. (7) Faunas evolve in isolation after barriers arise. (8) Vicariance is of primary significance in understanding modern patterns: related groups separated by barriers were fragmented by the appearance of the barriers.

J. Savage has presented a critique of vicariance theory and provides a dispersal-vicariance model that combines the best explanatory features of both approaches. *See* ECOLOGY. Jay M. Savage

Bibliography. H. G. Andrewartha and L. C. Birch, *The Ecological Web: More on the Distribution and Abundance of Animals*, 1986; J. C. Briggs, *Biogeography and Plate Tectonics*, 1987; J. H. Brown and A. C. Gibson, *Biogeography*, 2d ed., 1998; P. J. Darlington, *Zoogeography*, 1980; R. S. Dietz and J. C. Holden, The breakup of Pangaea, *Sci. Amer.*, 223(4):30–41, 1970; J. W. Hedgepeth, Classification of marine environments, *Mem. Geol. Soc. Amer.*, 67(1):17–27, 1957; G. Nelson and N. Platnik, *Systematics and Biogeography: Cladistics and Vicariance*, 1981; G. Nelson and D. E. Rosen, *Vicariance Biogeography*, 1981; J. M. Savage, The enigma of the Central American herpetofauna, *Ann. Missouri Bot. Gard.*, 69(3):444–556, 1983; D. S. Simberloff, Equilibrium theory of island biogeography and ecology, *Annu. Rev. Ecol. Syst.*, 5:161–182, 1974; G. G. Simpson, *The Geography of Evolution*, 1965.

Zoological nomenclature

The system of naming animals that was adopted by zoologists and detailed in the International Code of Zoological Nomenclature. The present system is founded on the 10th edition of C. Linnaeus's *Systema Naturae* (1758) and has evolved through international agreements culminating in the Code adopted in 1985. The primary objective of the Code is to promote the stability of the names of taxa (groups of organisms) by providing rules concerning name usage and the activity of naming new taxa. The rules are binding for taxa ranked at certain levels and nonbinding on taxa ranked at other levels. *See* ANIMAL SYSTEMATICS.

Zoological nomenclature is built around four basic features. (1) The correct names of certain taxa are either unique or unique combinations. (2) These names are formed and treated as Latin names and are universally applicable, regardless of the native language of the zoologist. (3) The Code for animals is separate and independent from similar codes for plants and bacteria. (4) No provisions of the Code are meant to restrict the intellectual freedom of individual scientists to pursue their own research.

There are four common reasons why nomenclature may change. (1) New species are found that were once considered parts of other species. (2) Taxonomic revisions may uncover older names or mistakes in identification of types. (3) Taxa may be combined, creating homonyms that require replacement. (4) Concepts of the relationships of animals change. Stability is subservient to progress in understanding animal diversity.

Code provisions. The 1985 Code applies to both living and extinct animals. It contains a series of interrelated articles governing the formation and use of names that form a very complex document. It specifies the conditions for determining whether a name is validly published, the publication date, and the authorship of the name. Validly published names are available for use. Of the several available names, the Code helps the taxonomist determine which is correct for a particular taxon: the Code uses the principle of priority that the oldest available name is the correct one unless unusual circumstances obtain. Coupled with the principle of priority is the concept of the first reviser, who is the first author to have treated a particular taxon after it was described. The decisions of the first reviser regarding names are usually followed except in unusual circumstances.

An International Commission on Zoological Nomenclature is empowered by the General Assembly of the International Union of Biological Sciences to suppress names that would normally be available if this is in the best interest of stability. The Commission also interprets the various articles and publishes opinions that are binding on all matters relating to the Code. The Code includes a number of appendices, such as a code of ethics and guides for transliteration and latinization of names.

The articles in the Code are directed toward the names of taxa at three levels. The family group includes taxa ranked as at the family and tribe levels (including super- and subfamilies). The genus group includes taxa below subtribe and above species. The species group includes taxa ranked as species or subspecies. Taxa above the family group level are not specifically treated, and their formation and use are not strictly regulated. For each group, provisions are made that are either binding or recommended.

Binominal nomenclature. The basis for naming animals is binominal nomenclature, that is, a system of two-part names. The first name of each species is formed from the generic name, and the second is a trivial name, or species epithet. The two names agree in gender unless the specific epithet is a patronym (named for a person). The combination must be unique; no other animal can have the same binominal. The formal name of a species also includes the author, so the formal name for humans is *Homo sapiens* Linnaeus. The genus, as a higher taxon, may have one or many species, each with a different epithet. *Homo* includes *H. sapiens, H. erectus, H. habilis*, and so forth. One feature of the Linnaean binominal system is that species epithets can be used over and over again, so long as they are used in different genera. *Tyrannosaurus rex* is a large dinosaur, and *Percina rex* is a small fresh-water fish. The epithet *rex* is not the species name of *Percina rex* because all species names are binominal in form. It is recommended that names of genera and species be set in a different typeface from normal text; italics is conventional. Different names

for the same species are termed synonyms, and the senior synonym is usually correct (principle of priority). If two species have the same name, they are considered homonyms, and the older name is correct. Homonyms are usually produced when two or more genera are combined, but homonymy is a complex subject and those interested are encouraged to consult the Code. Modern species descriptions are accompanied by a description that attempts to show how the species is different from others, and the designation of one or more type specimens. The type system is frequently viewed as pedantic "stamp collecting" by other biologists, but it is actually one of the foundations of nomenclature. The existence of actual specimens that can be examined by future generations provides physical evidence of what the author actually meant when describing the species.

Higher taxa. All higher taxonomic names have one part (uninominal) and are plural. Names of taxa of the family and genus groups must be unique. The names of genera are in Latin or latinized, are displayed in italics, and may be used alone. The names of the family group are formed by a root and an ending specific to a particular hierarchical level: family Hominidae (root + idae), subfamily Homininae (root + inae). The endings of superfamilies (root + oidea) and tribes (root + ini) are recommended but not mandated. The endings of taxa higher than the family group vary. For example, orders of fishes are formed by adding -iformes to the root (Salmoniformes), while in insects the ending is usually -ptera (Coleoptera). The principles of types, synonymy, and homonymy apply to the generic and family names, but not to higher taxa. *See* CLASSIFICATION, BIOLOGICAL; TAXONOMIC CATEGORIES. E. O. Wiley III

Bibliography. International Trust for Zoological Nomenclature, *International Code of Zoological Nomenclature adopted by the XX General Assembly of the International Union of Biological Sciences*, 1985; E. Mayr and P. D. Ashlock, *Principles of Systematic Zoology*, 2d ed., 1991; E. O. Wiley, *Phylogenetics, The Theory and Practice of Phyolgenetic Systematics*, 1981.

Zoology

The science that deals with knowledge of animal life. Together with botany, the science of plants, it forms biology, the science of living things. With the great growth of information about animals, zoology has been much subdivided. Some major fields are anatomy, which deals with gross and microscopic structure; physiology, with living processes in animals; embryology, with development of new individuals; genetics, with heredity and variation; parasitology, with animals living in or on others; natural history, with life and behavior in nature; ecology, with the relation of animals to their environments; evolution, with the origin and differentiation of animal life; and taxonomy, with the classification of animals. *See* ANATOMY, REGIONAL; DEVELOPMENTAL BIOLOGY; GENETICS; PARASITOLOGY; PHYLOGENY; PLANT EVOLUTION; TAXONOMY. Tracy I. Storer

Bibliography. C. E. Bond, *Biology of Fishes*, 2d ed., Saunders College Publishers, 1996; J. L. Gould and W. T. Keeton, *Biological Science*, 6th ed., W. W. Norton & Co., 1996; M. Hildebrand and G.E. Gaslow, Jr., *Analysis of Vertebrate Structure*, 5th ed., Wiley, 2001; K. V. Kardong, *Vertebrates: Comparative Anatomy, Function, Evolution*, 2d ed., McGraw-Hill, 1998; D. W. Linzey, *Vertebrate Biology*, McGraw-Hill, 2001; L. Sherwood, H. Klandorf, and P. H. Yancy, *Animal Physiology: From Genes to Organisms*, Thomson, Brooks/Cole, 2005; C. Starr and R. Taggart, *Biology: The Unity and Diversity of Life*, 10th ed., Thomson, Brooks/Cole, 2004; G. J. Tortora and B. Derrickson, *Principles of Anatomy and Physiology*, 11th ed., Wiley, 2006; T. A. Vaughan, J. M. Ryan, and N. J. Czaplewski, *Mammalogy*, 4th ed., Thomson, Brooks/Cole, 2000.

Zoomastigophorea

A class of protozoa of the subphylum Sarcomastigophora. Zoomastigophorea, also known as Zoomastigina, are flagellates which have few or no characters relating them to the pigmented forms. Some are simple, some are specialized; some have pseudopodia besides flagella, others have no pseudopodia. One group engulfs solid food at any body point; another shows localized ingestive areas. All are colorless. None produce starch or paramylum, and lipids and glycogen are assimilation products. Cells are naked or have delicate membranes. Colony formation is common. Colonies differ greatly in form and may be amorphous, linear, spherical, arboroid, or plane. Flagella vary from none to many. Nuclei generally have endosomes; frequently a rhizostyle extends from the nucleus to the flagellum base. Nutrition is holozoic, saprozoic, or parasitic, and encystment is common. Ecologically Zoomastigophorea are adapted to waters of relatively high organic content, but they occur in the clearest fresh waters and in the high seas. Nine orders are included in this class: the Choanoflagellida, Bicosoecida, Rhizomastigida, Kinetoplastida, Retortomonadida, Diplomonadida, Oxymonadida, Trichomonadida, and Hypermastigida. See articles on these orders. *See* PROTOZOA; SARCOMASTIGOPHORA. James B. Lackey

Zoonoses

Diseases that are transmitted from animals to humans. Zoonotic agents are found among all major categories of microbes, including bacteria, fungi, parasites, and viruses (see **table**). Although anyone can acquire a zoonotic infection, young children, older adults, and immunocompromised individuals are most at risk. Currently, there are large numbers of people that are considered immunocompromised

Some significant zoonoses*

Disease by category	Etiologic agent
BACTERIAL AND RICKETTSIAL	
Anthrax	*Bacillus anthracis*
Bordetella bronchiseptica pneumonia	*Bordetella bronchiseptica*
Brucellosis	*Brucella* sp.
Capnocytophaga	*Capnocytophaga canimorsus*
Cat-scratch, bacillary angiomatosis	*Bartonella henselae, Afipia felis*
Chlamydiosis/psittacosis	*Chlamydia psittaci*
Ehrlichiosis (monocytic and granulocytic)	*Ehrlichia chaffeensis, E. phagocytophilia*
Enteric disease	*Campylobacter jejuni, Clostridium perfringens, Escherichia coli, Salmonella* sp., *Vibrio parahemolyticus, Yersinia enterocolitica*
Leptospirosis	*Leptospira* sp.
Listeriosis	*Listeria monocytogenes*
Lyme disease	*Borrelia burgdorferi*
Murine typhus	*Rickettsia typhi*
Pasteurellosis	*Pasteurella multocida*
Plague	*Yersinia pestis*
Q fever	*Coxiella burnetti*
Rat bite fever	*Streptobacillus moniliformis*
Rickettsialpox	*Rickettsia akari*
Rhodococcosis	*Rhodococcus equi*
Scrub typhus	*Orienta tsutsugamushi*
Spotted fever, Rocky Mountain	*Rickettsia rickettsii*
Tuberculosis	*Mycobacterium bovis*
Tularemia	*Francisella tularensis*
MYCOTIC	
Cryptococcosis	*Cryptococcus neoformans*
Dermatomycosis or ringworm	*Microsporum canis, Trichophyton mentagrophytes*
PARASITIC	
African trypanosomiasis, sleeping sickness	*Trypanosoma brucei rhodesiense*
American trypanosomiasis or Chagas' disease	*Trypanosoma cruzi*
Anisakiasis	*Anisakis* sp., *Psuedoterranova* sp., *Contracaecum* sp.
Balantidiasis	*Balantidium coli*
Clonorchiasis	*Clonorchis sinensis*
Coccidiosis	*Cryptosporidum parvum, Sarcocystis* sp.
Cutaneous larva migrans	*Ancylostoma canium, A. braziliense*
Diphyllobothriasis or fish tapeworm	*Diphyllobothrium latum*
Dipylidiasis or dog tapeworm	*Dipylidium caninum*
Dirofilariasis	*Dirofilaria immitus, D. repens*
Giardiasis	*Giardia lamblia*
Hydatid disease or echinococcosis	*Echinococcus* sp.
Paragonimiasis	*Paragonimus westermani*
Leishmaniasis	*Leishmania donovani, L. tropica,* and other species
Piroplasmosis, babesiosis	*Babesia microti* and other species
Taeniasis or cysticercosis	*Taenia saginatum, T. solium*
Toxoplasmosis	*Toxoplasma gondii*
Trichinosis	*Trichinella* sp.
Visceral larva migrans	*Toxocara canis, T. cati, Baylisasearis procyonis*
VIRAL	
Ebola and Marburg hemorrhagic fevers	RNA Filovirus
Encephalitis (La Cross, California)	RNA Bunyavirus
Equine encephalitis (Eastern, Western, Venezuelan)	RNA Togavirus
Hantavirus pulmonary syndrome	RNA Hantavirus
Hemorrhagic fevers: Argentine, Bolivian, Brazilian, Venezuelan	RNA Arenavirus
Hemorrhagic fever: Crimean-Congo	RNA Nairovirus
Hemorrhagic fever: renal	RNA Hantavirus
Hendra or Nipah diseases	RNA Paramyxovirus
Hepatitis	RNA Hepatitis A
Herpes B	DNA Herpes simiae
Lassa fever	RNA Arenavirus
Lymphocytic choriomeningitis	RNA Arenavirus
Monkey pox, cowpox	DNA Orthopoxviruses
Orf	DNA Parapoxvirus
Rabies	RNA Lyssavirus
Rift Valley fever	RNA Bunyavirus
Vesicular stomatitis	RNA Vesiculovirus
West Nile fever	RNA Flavivirus

*Based on data from William T. Hubbert, Ann Marie Nelson, and Brett Saladino.

(for example, people receiving chemotherapy or those infected with agents such as HIV).

Animals can also serve as reservoirs for zoonotic agents; that is, they are a source of the microorganism that allows transmission to the human population. Domesticated animals can often serve as a bridge for zoonotic agents that are commonly found in feral animal populations. However, sometimes animals serve only as indicators that an organism is present in the environment. In these cases, they become infected from contaminated soil, water, or air, as do humans. Additionally, there are diseases, called anthroponoses, that are transmitted from humans to animals.

Nervous system infections. Important zoonotic agents that can invade the nervous system in humans include *Toxoplasma gondii* and *Cryptococcus neoformans.*

Toxoplasma. Thirty to forty percent of adult humans in the United States are seropositive for *T. gondii*. Feline species are the only definitive hosts for the organism. Cats harbor sexual stages of the parasite in their gastrointestinal tract and shed infectious oocysts in their stool. The organism has been found in all areas of the world, in over 200 species of birds, and in many species of warm-blooded animals, which act as intermediate hosts. Intermediate hosts become infected by ingesting sporulated oocysts from cats. Humans can contract the disease by ingesting inadequately cooked meat of the intermediate hosts with infectious bradyzoites and tachyzoites, or by ingesting oocysts from soil contaminated with cat feces. In one European study, more than 50% of the meat of pigs and sheep carried toxoplasmosis. Humans can be infected by ingesting any source of meat with *Toxoplasma*, but lamb, mutton, and pork are most commonly infected. *See* TOXOPLASMIDA.

Cryptococcus. Cryptococcus neoformans is an important pathogen in immunosuppressed patients and occurs commonly as a premonitory sign of acquired immune deficiency syndrome (AIDS), where it causes severe life-threatening meningitis. *Cryptococcus* is taken up by the human host via the respiratory tract. The organism has been found in soil, plants, bird feces, raw milk, fruit juices, and from the oral cavity, gastrointestinal tract, and skin of healthy humans. It is commonly found in pigeon roosts and in soil contaminated by pigeon feces, and has been isolated from droppings of other bird species. *Cryptococcus neoformans* has an ecological relationship with pigeons in that it selectively assimilates creatinine in pigeon urine. *See* ACQUIRED IMMUNE DEFICIENCY SYNDROME (AIDS).

Respiratory tract infections. *Bordetella bronchiseptica* is a secondary bacterial invader that occurs commonly in the respiratory tracts of dogs, cats, and pigs, and less commonly in horses. It complicates the management of viral respiratory disease in the dog, and has been associated with clinical signs of fever, anorexia, coughing, and nasal discharge. Immunocompromised patients should avoid exposing themselves and their dogs to environments with sources of *B. bronchiseptica* such as dog shows, kennels, or other settings where dogs are housed closely together.

Gastrointestinal tract infections. Agents causing primary gastrointestinal disease that have been documented to be transmitted to humans from an animal source include *Salmonella, Campylobacter, Mycobacteria, Cryptosporidia*, and *Giardia*. Agents that have been implicated as zoonotic agents include *Entamoeba, Isospora, and Microsporidia*.

Salmonella. Salmonella is an important pathogen in humans. Salmonellosis is a major communicable disease problem in humans in the United States, with the infection rate estimated as high as 2 million cases per year. Poultry and beef products are the largest sources of nontyphoid *Salmonella* in the United States. Contaminated animals and animal products are a major source of *Salmonella* for humans. The most common source is from contaminated foods. Strict hygiene in food preparation is needed to avoid contaminating cold-served food products (such as salads and coleslaw) with raw meat. It is estimated that the rate of infection of *Salmonella* in domestic animals is 1–3%. Transmission of antibiotic-resistant *S. typhimurium* to humans has been reported due to contact with sick calves shedding the organism.

Salmonellosis is frequently diagnosed in dogs and cats, and these animals can also be a source of infection for humans. Confining dogs and cats in the house as much as possible minimizes their contact with fecal material of other animals. Avoidance of feeding pets raw meat prevents their contracting the disease from a well-known source of salmonellosis. Animals with diarrhea should be presented promptly to a veterinarian for diagnosis. Treatment of pets with salmonellosis is usually not recommended unless clinical signs of bacteremia are evident, because treatment may prolong the period of shedding of the organism. Pets should not be returned to the household of immunosuppressed individuals until two fecal cultures taken 24 hours apart are negative for *Salmonella*. It is important that these fecal cultures be taken after any antibiotic treatment has ended. *Salmonella* is also carried commonly by cold-blooded animals. *See* SALMONELLOSES.

Campylobacter. Campylobacter is another common gastrointestinal and blood infection that can be transmitted from animals to humans. *Campylobacter* is a significant cause of enteritis in humans all over the world; it has been targeted as one of four bacterial pathogens for elimination from human food supplies by United States agencies. *Campylobacter jejuni* is the most important species, although other species of *Campylobacter* can cause diarrhea in humans. Transmission of *C. jejuni* to humans occurs commonly from contaminated food and water. Meat products, especially poultry, are frequently contaminated with this bacteria. *Campylobacter jejuni* is found generally in water sources, where it can live for several weeks at low temperature. Many animals, including cattle, sheep, swine, fowl, dogs, and cats, excrete *Campylobacter* spp. Household pets, especially kittens and puppies, frequently shed these organisms, with an incidence of isolation of

Campylobacter in feces as high as 35–42%. To prevent infection with *Campylobacter*, people should cook meat products thoroughly and wash cooking utensils well. Immunocompromised individuals should be aware of the risks of owning or coming in contact with puppies and kittens and their excreta. Animals with diarrhea should be presented promptly to their veterinarian, and a fecal sample should be taken for testing for *Campylobacter*. If *Campylobacter* spp. are isolated from dogs or cats, treatment with erythromycin is recommended to eliminate shedding of the organism in feces.

Cryptosporidium. Cryptosporidium is a protozoan parasite that is only 1–6 micrometers in size. It lacks host specificity and infects the gastrointestinal tract of a wide range of mammals.

Cryptosporidium parvum is the most important species in humans and domesticated animals. It has been shown to occur naturally in mice, rabbits, guinea pigs, cats, dogs, squirrels, raccoons, horses, pigs, sheep, goats, cattle, and several species of nondomesticated ruminants. Isolates from humans have proven infective to cats, cattle, dogs, goats, mice, pigs, rats, sheep, and other humans. Transmission of *C. parvum* to humans from cats, cattle, and pigs has also been reported. *Cryptosporidium parvum* is a common cause of diarrhea in young calves who acquire the organism from their mothers and the environment; it is commonly shed in the feces of calves with no clinical signs of disease. Fecal contamination of food and water contribute to the high prevalence of cryptosporidial infection in persons residing in areas with poor sanitary conditions. Contaminated food, fomites, flies, and water are sources of infection with *Cryptosporidium*. Oocysts have proven resistant to most chemical disinfection at approved concentrations, and disinfection of drinking water by common methods is not effective.

Giardia. Giardia is a flagellate protozoan parasite that infects the gastrointestinal tract of many species of animals including dogs, cats, horses, pigs, cattle, sheep, goats, and many other mammals. It naturally infects humans, and is endemic in many developing countries where the prevalence in children is as high as 30%. Pets are potential sources of *Giardia* for humans. Studies in New Jersey have shown infection rates of 35% and 2.5% in stray dogs and stray cats, respectively. Other studies found high infection rates in cattle, beavers, and coyotes. Treatment for giardiasis is difficult in companion animals. Infected animals should be isolated from other animals in the household, and treated and bathed to remove cysts from the fur. Strict personal and food hygiene minimizes the risk of *Giardia* infection from human sources. *See* GIARDIASIS.

Parasitic diseases of animals. Some parasitic diseases of animals have zoonotic potential. The dog heartworm, *Dirofilaria immitis*, is endemic in many areas of North America and the rest of the world, requiring daily chemical prophylaxis in pet dogs during the mosquito season. Humans can be accidentally infected by *D. immitis* microfilaria-carrying mosquito vectors due to the high prevalence of *D. immitis*, although the disease is rare. The disease in humans is often diagnosed when a solitary lung lesion or coin lesion is noted on a chest radiograph and confirmed by biopsy.

Toxocara. Toxocara canis and *T. cati* are round worms of the dog and cat. Humans acquire infection with these parasites by ingesting embryonated eggs shed in the feces of animals. Visceral larva migrans occurs when eggs hatch, penetrate the intestine, and migrate through body tissue. Migration can result in multiple abscesses, eosinophilic granulomas, hepatomegaly, and pneumonitis with eosinophilia. Symptoms include coughing, nausea, vomiting, and dyspnea (difficulty in breathing). Larva migrans can also affect the eyes. Ocular larva migrans causes inflammation, fibrosis, and loss of vision. The disease most often occurs in children less than 3 years of age, but can also occur in adults. The infant and toddler age group is most predisposed due to the tendency to ingest soil. Eggs of *Toxocara* are extremely resistant in soil and remain infectious for years. Routine yearly fecal exams and treatment of positive animals with anthelminthics is recommended to decrease the potential for this problem.

Ancylostoma caninum and *A. brasiliense* are hookworms of dogs and cats which rarely cause cutaneous larva migrans in humans. Hookworm eggs are shed in feces into the environment, hatch, and undergo two larval molts to produce infectious, third-stage larvae. Hookworms infect their definitive hosts by penetrating the skin. In the correct host, larvae migrate from the skin to the lungs, ascend the respiratory tree, and are swallowed to resume development in the intestine where they mature into adult hook worms. Accidental infection of animal hookworms in humans usually occurs through larval penetration of the extremity when walking barefoot through infected soil. After skin penetration, the larvae migrate in the germinal layer of the epidermis, causing sinuous tunnels with associated itching and inflammatory response. The larvae can survive in the skin for several weeks or months. Although some human infections may be associated with pneumonitis and eosinophilia when the larvae invade the lung or other organ tissue, most cases are limited to the skin.

Echinococcus. Echinococcus granulosus and *E. multilocularis* are tapeworms that cause hydatid disease or echinococcosis. *Echinococcus granulosus* is a tapeworm whose life cycle involves dogs, coyotes, wolves, and dingos as the primary hosts and sheep, swine, cattle, moose, caribou, and humans as intermediate hosts. This parasite is endemic in North and South America, England, Africa, the Middle East, Australia, and New Zealand. In the United States, it occurs mainly in the western states. *Echinococcus multilocularis* is a tapeworm that cycles between dog, cat, and fox as definitive hosts, and between vole, lemming, cattle, horse, swine, and humans as intermediate hosts. *Echinococcus multilocularis* is endemic in north-central Europe, Alaska, Canada, and the central United States far south as Illinois and Nebraska. Dogs and cats that ingest wild game and rodents can be infected with

E. multilocularis and can be a risk to humans. Humans act as intermediate hosts for both of these parasites. Humans ingest *Echinococcus* eggs from feces of the definitive hosts, which hatch in the intestines releasing a larval tapeworm that migrates and develops in tissues. Hydatid cysts are formed which are found in the liver, lung, nervous system, bone, spleen, kidney, and heart. Symptoms range from inapparent to severe and relate to the site of the cyst and whether it ruptures. Prevention of infection involves avoidance of contact with fecal matter from pets and wild carnivores, good personal and food hygiene, keeping pets indoors, and routine fecal exams of pets for parasites.

Fleas. Fleas are common, difficult to eliminate, and capable of transmitting several important zoonotic diseases. However, in the United States fleas are mostly a nuisance problem causing pruritis and hypersensitivity in animals and humans. Control of fleas must focus on the environment of the pet, such as the house and the yard. Pets may be treated for fleas weekly, but treatment must coincide with treatment of the environment.

Dipylidium caninum is a common tapeworm of dogs and cats. Cats and dogs are infected during grooming by ingesting the flea intermediate host containing an infective cysticercoid. Cysticercoids mature into adult tapeworms in the intestine of their normal host. Tapeworm eggs that are shed in the feces of an infected dog or cat are eaten by flea larvae, where they hatch and develop to the cysticercoid in the body cavity of the flea. Humans, especially children, are accidentally infected when they swallow a flea while crawling on the floor or while playing with pets.

Ticks. Animals are exposed to ticks when they roam fields and wooded lots. Ticks transmit many zoonotic diseases, including babesiosis, tick fever, rickettsial diseases, Lyme disease, Q-fever, relapsing fever, typhus, ehrlichiosis, and tularemia. Humans are exposed to ticks by similar means, but in some cases can acquire ticks from animals. Pets kept indoors are less likely to acquire ticks. Humans can use tick repellants, check themselves for ticks after a walk in the woods, and avoid tick-infested areas. *See* LYME DISEASE; RICKETTSIOSES; TULAREMIA.

Systemic disease. *Bartonella* spp., including *B. henselae* (formerly *Rochalimaea*), a member of the rickettsia family, have been implicated as causative agents in cat scratch disease and bacillary angiomatosis (vascular tumors of the skin, internal organs, bone marrow, and lymph nodes). In humans, cat scratch disease is characterized by persistent regional lymphadenopathy in the lymphatic drainage area of the site of a recent cat bite or cat scratch. In immunocompetent hosts, catch scratch disease tends to be a disease of children, usually self-limited and localized to regional lymph nodes. It rarely causes systemic or widespread infections. In HIV-infected hosts, the disease usually affects young adults and manifests itself as bacillary angiomatosis. Manifestation of bacillary angiomatosis includes cutaneous nodules resembling Kaposi sarcoma, encephalopathy (disease of the brain), lymphadenopathy, pneumonitis (inflammation of lung tissue), bone lesions, hepatosplenomegaly (lenlargement of the liver and the spleen), and hypotension (low blood pressure) with metabolic acidosis. The disease in HIV-infected individuals is widespread, but responds to antibiotic therapy. *Bartonella henselae* is widespread in the cat population, with isolation from the blood of 41–46% of cats in studies involving household, stray, and pound cats. Since arthropods are known vectors for the transmission of rickettsias, prevention should be focused on flea control and minimizing direct transmission by cats. Specific measures include declawing of pet cats, avoiding bites or scratches from cats, and strict flea control. *See* BARTONELLOSIS; CAT SCRATCH DISEASE.

Pasteurella multocida is an important pathogen in animal bites and a documented life-threatening pathogen in immunocompromised patients. *Pasteurella multocida* has been found in the oropharynx of 60–90% of healthy cats and in 55% of dogs. Transmission occurs via traumatic wounds such as bites and scratches by the respiratory route via aerosol, and by ingestion. Dissemination and septicemia (persistence and multiplication of living bacteria in the blood) have been noted in patients on chemotherapy, with cirrhosis, or with hematologic malignancy. Two cases of *P. multocida* have been documented in an HIV-infected individual. One patient undergoing peritoneal dialysis developed peritonitis, and another patient developed sinusitis and pneumonia apparently from aerosol transmission by a cat. To minimize transmission of *P. multocida*, one should avoid being bitten or scratched by pet dogs and cats. *See* PASTEURELLA.

There are many sources of information on zoonotic diseases, including nonprofit organizations and the United States Centers for Disease Control. Overall, the risk of contracting zoonotic diseases by susceptible individuals is likely to be small, but physicians and veterinarians can decrease this risk by educating patients on the transmission and prevention of these diseases. *See* ANIMAL VIRUS; MEDICAL BACTERIOLOGY; MEDICAL PARASITOLOGY.

Linda S. Mansfield

Bibliography. F. J. Angulo et al., Caring for pets of immunocompromised persons, *J. Amer. Vet. Med. Ass.*, 205(12):1711–1718, 1994; E. U. Canning, Cryptosporidia, in J. P. Kreier (ed.), *Parasitic Protozoa*, Academic Press, New York, 1993; Centers for Disease Control and Prevention, *USPHS/IDSA Guidelines for the Prevention of Opportunistic Infections in Persons Infected with Humans Immunodeficiency Virus: A Summary*, MMWR 44(No. RR–8):1–34, 1995; J. P. Dubey, Toxoplasmosis, *J. Amer. Vet. Med. Ass.*, 205(11):1593–1598, 1994; I. Nachamkin, M. J. Blaser, and L. S. Tompkins, *Campylobacter jejuni: Current Status and Future Trends*, American Society for Microbiology, Washington, DC, 1992; C. R. Sterling and M. J. Arrowood, Cryptosporidia, in J. P. Kreier (ed.), *Parasitic Protozoa*, Academic Press,

New York, 1993; P. G. Wall et al., Transmission of multi-resistant strains of *Salmonella typhimurium* from cattle to man, *Vet. Rec.*, 136:591–592, 1995.

Zooplankton

Animals that inhabit the water column of oceans and lakes and lack the means to counteract transport currents. Zooplankton inhabit all layers of these water bodies to the greatest depths sampled, and constitute a major link between primary production and higher trophic levels in aquatic ecosystems. Many zooplankton are capable of strong swimming movements and may migrate vertically from tens to hundreds of meters; others have limited mobility and depend more on water turbulence to stay afloat. All zooplankton, however, lack the ability to maintain their position against the movement of large water masses.

Zooplankton can be divided into various operational categories. One means of classification is based on developmental stages and divides animals into meroplankton and holoplankton. Meroplanktonic forms spend only part of their life cycles as plankton and include larvae of benthic worms, mollusks, crustaceans, echinoderms, coral, and even insects, as well as the eggs and larvae of many fishes. Holoplankton spend essentially their whole existence in the water column. Examples are chaetognaths, pteropods, larvaceans, siphonophores, and many copepods.

Size is another basis of grouping all plankton. A commonly accepted size classification scheme includes the groupings: picoplankton (<2 micrometers), nanoplankton (2–20 μm), microplankton (20–200 μm), mesoplankton (0.2–20 mm), macroplankton (20–200 mm), and megaplankton (>200 mm).

Systematic composition. Nearly every major taxonomic group of animals has either meroplanktonic or holoplanktonic members. Some of the more common zooplankton groups (see **illustration**) are described below.

Protists are single-celled eukaryotes that are vital to planktonic systems. Photosynthetic protists form the base of planktonic food webs as the major primary producers in the plankton. Animallike protists (protozoa) constitute the largest proportion of the nano- and microzooplankton and include amebas, nonphotosynthetic flagellates, and ciliates. Some of the ameboid Foraminifera and Actinopoda with tests (rigid skeletal supports) or hard skeletons are larger and reach sizes exceeding 2 mm. The skeletons of foraminiferans and actinopods form an important part of the deep-sea sediments, and foraminiferans are used as markers in oil exploration. Nanoplanktonic flagellates are usually the most important predators of bacteria; however, some dinoflagellates can capture and digest larger organisms using protoplasmic nets. Both photosynthetic and nonphotosynthetic dinoflagellates have been implicated in noxious red tides or fish kills. Various ciliates species capture bacteria, phytoplankton, and other protists. Many protists that do not fall neatly into a classification as animallike (such as protozoa) or plantlike organisms are termed mixotrophs. Mixotrophs bridge the plant/animal division because they have both photosynthetic and phagotrophic capabilities. Some flagellates with chloroplasts ingest bacteria, phytoplankton, or other protists. Conversely, some ciliates retain functional chloroplasts from phytoplankton that are ingested and otherwise digested. *See* FORAMINIFERIDA; PROTOZOA; RADIOLARIA.

The phylum Cnidaria contains a number of groups which are important in the marine plankton, including scyphozoans, the true jellyfish. Both scyphozoans and the colonial siphonophores are carnivores with tentacles bearing stinging nematocyst cells. They are found at all depths in the ocean but are most common in the upper waters. Cnidaria are rare in fresh water.

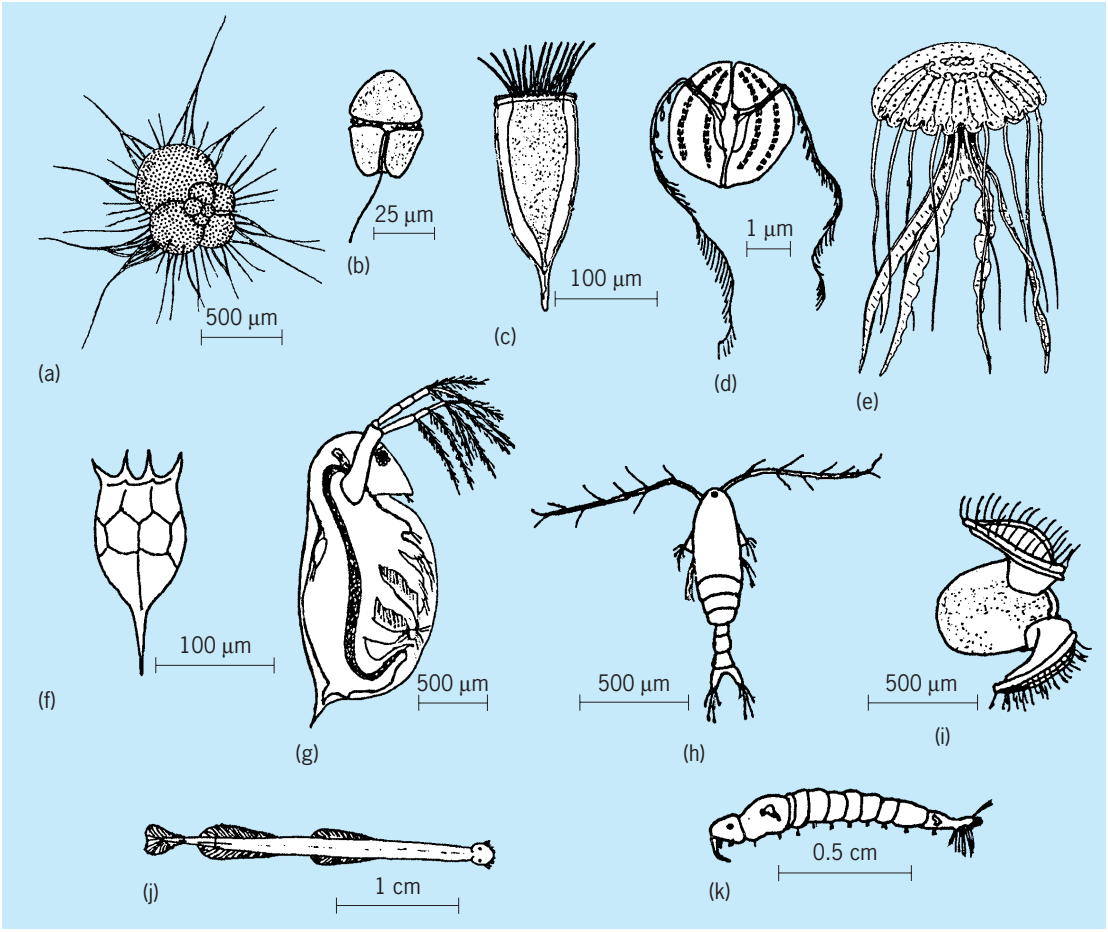
The ctenophores, or comb jellies, were once grouped with Cnidaria but are now treated separately because they lack nematocysts. All are exclusively carnivorous and are important predators of many zooplankton, especially copepods. *See* COELENTERATA.

Rotifers probably arose in fresh water; only about 5% of the approximately 2500 described species are found in marine and brackish waters. Although about a hundred species are holoplanktonic, most are primarily sessile. Nevertheless, they are often a numerically important fraction of the zooplankton. Most rotifers feed on bacteria, detritus, and algae, although some are predators of protozoa and other rotifers. Under favorable conditions, rotifers exhibit parthenogenesis (egg production without fertilization leading to female-only populations). Sexual reproduction does occur, but only after male eggs are produced as a response to environmental stresses. Cyclomorphosis (seasonal changes in morphology within a species) is common among rotifers. *See* ROTIFERA.

Chaetognatha is a carnivorous marine group with worldwide distribution. They are mostly holoplanktonic and are found at all depths, although a given species may be restricted to certain water masses and depths. The elongate body is transparent with lateral fins and is usually less than 3 cm in length. *See* CHAETOGNATHA.

Veliger larvae of many benthic mollusks are frequently seen in coastal plankton. There are also marine gastropods, including heteropods and pteropods, which are adapted to a holoplanktonic life style. Another group represented in shallow marine waters by larvae is Polychaeta. This annelid class also contains a few holoplanktonic families. *See* MOLLUSCA; POLYCHAETA.

Copepods are almost always the most numerous members of the macrozooplankton community in marine systems, and are important in freshwater as well. They often migrate to deeper water in daytime and move toward the surface at night, using the first antennae and the thoracic appendages to



Representative zooplankton. (*a*) Foraminifera. (*b*) Dinoflagellate. (*c*) Tintinnid ciliate. (*d*) Ctenophore. (*e*) Cnidarian (scyphozoan jellyfish). (*f*) Rotifer. (*g*) Cladoceran. (*h*) Copepod. (*i*) Gastropod veliger larva. (*j*) Chaetognath. (*k*) Insect larva (*Chaoborus*). (*Parts a and e from L. H. Hyman, The Invertebrates, vol. 3: Protozoa through Ctenophora, McGraw-Hill, 1940*)

swim. Some are primarily suspension feeders and use appendages with hairlike setules to remove phytoplankton and detritus from the water. Others are primarily raptorial and feed on larger particles, including smaller zooplankton. *See* COPEPODA.

Cladocera are a second major group of planktonic crustaceans. They often outnumber copepods in fresh water and occasionally are abundant in coastal marine waters. Most have bodies covered by a folded carapace that gives an appearance of being bivalved. The enlarged second antennae are the primary swimming appendages. For most Cladocera the primary food source is phytoplankton, which are filtered from water passing through appendages within the carapace. A few species are predaceous on other zooplankton. Like copepods, cladocerans are well known for diel (diurnal) vertical migrations. Their occurrence tends to be seasonal, with a resting egg formed between periods of rapid parthenogenic reproduction. Cyclomorphosis is common.

Euphausids or krill, are relatively large marine crustaccans (up to 3cm) that are found worldwide at all depths. Individual species have more limited distribution. Some species, especially cold-water forms, demonstrate swarming behavior, and in high latitudes they are the major food source of baleen whales. The most important food of polar euphausids appears to be diatoms, while carnivorous species are more common in warmer waters. *See* EUPHAUSIACEA.

Most aquatic insect species tend to be associated with the bottoms, shores, or surface layers of streams and lakes. The only major insect member of the plankton is the larvae of the phantom midge *Chaoborus*. This larvae is able to resist anoxia (lack of oxygen) and is often found at the bottom during day and in the plankton at night. *Chaoborus* species are important predators of other zooplankton in lakes.

Tunicates are all marine. Many are benthic and produce planktonic larvae, but two classes, Appendicularia and Thaliasia, are holoplanktonic. Appendicularia are neotenic and build a mucous house which acts as a filter to capture small food particles. The Thaliacea, including salps and doliolids, are also filter feeders. *See* TUNICATA.

Adaptations. A problem faced by all plankton is to maintain position in the water column. Flattened bodies and numerous lateral spines or plumose setae which increase surface-to-volume ratios are common in various zooplankters. This increases resistance to the passage of water and thus slows

sinking. Other adaptations to a floating existence are positive buoyancy mechanisms, such as oil droplets, gas-filled floats, or regulation of ionic balance by replacement of heavy ions with lighter ones. Some zooplankton have gelatinous sheaths that appear to slow the sinking rate. For example, the gelatinous sheath produced by the cladoceran *Holopedium gibberum* reduces its sinking rate by about 50%.

Life in open water exposes zooplankton to heavy predation by visual predators, especially fishes. Many zooplankters are nearly transparent, which affords them some protection. Others, such as fresh-water mites, have conspicuous coloration to advertise their noxious taste, so that visual predators learn to avoid them. In the blue light that penetrates to deep water, reddish hues appear black and thus invisible; it is not surprising that many bathypelagic zooplankton are red. Another antipredation adaptation is cyclomorphosis. Kairomones (chemical messengers that are beneficial to the recepient) released by predators appear to induce cyclomorphosis in succeeding generations of rotifers and cladocerans. These changes in morphology, such as increased spine lengths or head shields, reduce predation. Some adaptations have multiple advantages; the gelatinous sheath of *Holopedium* reduces predation by invertebrates in addition to slowing the sinking rate. *See* ADAPTATION (BIOLOGY).

Vertical migration. Most major zooplankton groups have at least some species that display diel migrations which usually consist of downward movement during the day and upward movement at night. The distance traveled can be hundreds of meters. The same species may display the classical day-down, night-up movement in one area and display no migratory behavior in another location. The sex and age of the zooplankton, as well as the season, can affect their vertical position in the water column and the degree of migration observed.

Diurnal light variation is the most likely mechanism triggering vertical migration. Many planktonic animals are positively phototactic at low light intensities and negatively phototactic at high intensities. Although the general consensus is that light is a major stimulus for the timing of migration, there are many explanations for its purpose. Zooplankton may sink to depths where illumination is insufficient for detection by visual predators in daytime, while at night, when visual predators do not hunt, zooplankton can return to the surface to feed on phytoplankton. Another explanation is that zooplankton remain in deeper, colder water during the day to reduce their metabolism, and return to the surface at night to feed on phytoplankton. The energetic advantage gained by remaining in colder water must exceed that expended in migration and lost due to lack of continuous feeding at the surface. Additionally, the potentially reduced food quantity and the cold temperature of the deeper water result in slower egg development and thus lower reproductive rates. Differential migration by ecologically similar species could also reduce competition for resources. It is unlikely that a single factor can explain all vertical migration; the above factors and others probably interact, or are important at different times.

Communities. Zooplankton, like all organisms, have a range of environmental conditions to which they are adapted. The optimum environment for one species may be barely tolerable to another. Physical and chemical boundary conditions, including turbulence, light, temperature, and salinity gradients, are important in determining species makeup of a zooplankton community. For example, plankton on the two sides of the Gulf Stream or within the upper and lower waters of a lake differ considerably. Some zooplankton have such clear-cut ecological demands that the presence of particular species can indicate the origin of the water mass. Narrow temperature tolerances limit some taxa to tropical or polar waters or to certain periods of the year in temperate zones. Differential salinity tolerances are reflected by changes in the composition of a community as estuaries become increasingly brackish downstream. Thus zooplankton populations are not distributed homogeneously, but tend toward both vertical and horizontal patchiness.

In addition to the physical and chemical boundary conditions mentioned above, biological interactions such as predation, food availability, reproduction, and social behavior affect the distribution of zooplankton and the resulting community structure. Trophic interactions, that is, "what eats what," are of major importance in structuring zooplankton communities. The presence of fishes that feed selectively on larger zooplankton can limit the species composition to smaller-bodied animals. Lakes without size-selective vertebrate predators usually have a higher proportion of larger invertebrate plankton. Likewise, selective feeding by zooplankton on various algae and protozoa will alter the community structure of these groups.

Seasonal breeding of holoplanktonic organisms and entrance of larval meroplankton into the water column play a large part in changing community composition. Food supply, temperature, and other factors interact to determine seasonal breeding patterns in zooplankton. Increasing light and temperature in temperate and boreal areas leads to spring phytoplankton blooms, when many zooplankters release their young. Other groups have maximum abundances in summer or during a secondary peak of primary production in the fall. *See* ECOLOGICAL COMMUNITIES.

Productivity. Productivity is the amount of biomass generated per unit time. There is extensive geographic variation in the pattern and magnitude of productivity of the phytoplankton that are the base of planktonic food webs. This variation in primary production dictates a similar difference in the secondary productivity of zooplankton in both lakes and marine systems. Zooplankton tend to grow faster or produce more young when high-quality food is readily available. However, organisms are not 100% efficient at converting food to biomass. Some

ingested food is not assimilated into the animals' bodies, and some assimilated food is lost to maintenance metabolism and respiration; only what is left becomes available for secondary production (growth and reproductive output). Although conversion efficiencies as high as 90% have been determined for individual zooplankton species feeding on high-quality foods in laboratory experiments, 10–30% efficiencies are more common. Zooplankton production is difficult to determine for natural populations because of the diversity of zooplankton species, the differential digestibility and availability of their food sources, and temperature-induced alterations in their metabolism, among other reasons. *See* BIOMASS.

Planktonic food web. The classic description of the trophic dynamics of plankton is a food chain consisting of algae grazed by crustacean zooplankton which are in turn ingested by fishes. This model may hold true to a degree in some environments such as upwelling areas, but it masks the complexity of most natural food webs. Zooplankton have an essential role in linking trophic levels, but several intermediate zooplankton consumers can exist between the primary producers (phytoplankton) and fish. Thus, food webs with multiple links to different organisms indicate the versatility of food choice and energy transfer and are a more realistic description of the planktonic trophic interactions.

Size is of major importance in planktonic food webs. Most zooplankton tend to feed on organisms that have a body size smaller than their own. However, factors other than size also modify feeding interactions. Within a range of ingestible sizes, dependent on species and age, zooplankton can track changes in the particle-size spectrum of phytoplankton and graze the most abundant size classes. Some phytoplankton are noxious and are avoided by zooplankton, and others are ingested but not digested. Furthermore, zooplankton frequently assume different feeding habits as they grow from larval to adult form. They may ingest bacteria or phytoplankton at one stage of their life cycle and become raptorial feeders later. Other zooplankton are primarily herbivorous but also ingest heterotrophic protists and can opportunistically become carnivorous. Consequently, omnivory, which is considered rare in terrestrial systems, is a relatively common trophic strategy in the plankton. In all food webs, some individuals die without being consumed and are utilized by scavengers and ultimately by decomposers (bacteria and fungi). *See* ECOLOGY; ECOSYSTEM; MARINE ECOLOGY; PHYTOPLANKTON.

Robert W. Sanders

Bibliography. W. Lampert and U. Sommer, *Limnoecology*, 1997; M. Omori and I. Tsutomu, *Methods in Marine Zooplankton Ecology*, 1992; U. Sommer (ed.), *Plankton Ecology*, 1989; T. Tamminen and H. Kuosa (eds.), *Eutrophication in Planktonic Ecosystems: Food Web Dynamics and Elemental Cycling*, 19XX; J. Thorp and A. Covich (eds.), *Ecology and Classification of North American Freshwater Invertebrates*, 1991; C. D. Todd, M. S. Laverack, and G. Boxshall, *Coastal Marine Zooplankton*, 1996.

Zoraptera

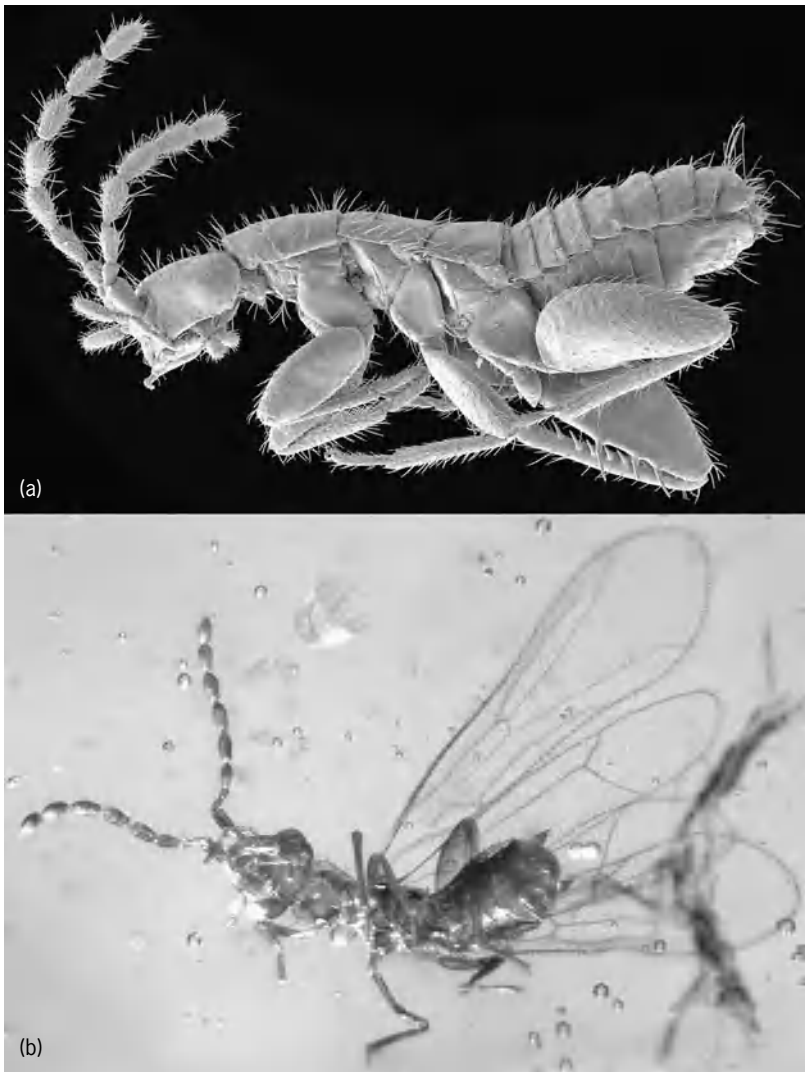
An order of minute insects superficially resembling termites (Isoptera) or booklice (Psocoptera). Although species nest in decaying wood, they are of no economic importance. Interest in the group principally derives from their unique biology, anatomy, and evolutionary relationship to other insects.

Morphology and development. Zorapterans are generally less than 3 mm (0.11 in.) in total body length and can be distinguished by their characteristic two-segmented tarsi; unsegmented cerci; nine-segmented antennae (which is reduced to eight segments in two fossil species from the Cretaceous); dehiscent, paddle-shaped wings with uniquely reduced pattern of venation; greatly enlarged hind femora bearing rows of stiff spines along their undersurfaces; asymmetrical male genitalia; and vestigial ovipositor. Individuals of each species occur in two morphs (sometimes referred to as "castes" but not to be confused with the actual castes of truly social insects): eyed and winged forms that shed their wings after dispersal (becoming what are called "dealates"), or blind and wingless forms that predominate in colonies (see **illustration**).

Development progresses through a series of nymphal stages. The relatively soft integument of zoopterans is pale in nymphal stages and typically reddish brown in adults.

Ecology. Zoraptera are gregarious, living in small colonies of 15–120 individuals in crevices or under bark of moist, decaying logs. Zorapterans will sometimes settle in human-made sawdust piles, but such colonies tend not to last as long as those in logs or stumps. Species prefer rotting wood that has decomposed to the point that logs can be easily torn through by hand or with an ordinary garden tool. Colonies harbor in naturally formed spaces in logs that are not reached by light. Once the wood is disturbed, individuals quickly scatter to avoid detection. Species feed principally on fungal hyphae and spores but can also be generalist scavengers or predators, victimizing nematodes, mites, or other tiny arthropods.

Distribution. Zoraptera are principally distributed pantropically, with only four species occurring north of the Tropic of Cancer (that is, North of 23.5°N): two species in North America and two in Tibet. Fossil records for Zoraptera are all from ambers formed in warm, tropical paleoclimates. Although species have at times been considered highly endemic, due to their poor dispersal abilities, they are increasingly being discovered to have larger geographic ranges than previously understood and are apparently capable dispersers. The presence of zorapterans on distantly isolated islands of relatively recent geological age attests to their dispersal capabilities. In fact, the "rarity" of Zoraptera may be more the result of poor collecting than any actual scarcity. While some species are assuredly uncommon, the actual abundance of zorapterans does not live up to its reputation for scant occurrence.



Representative zorapterans. The greatest length of both individuals is 2.5 mm. (*a*) Adult female of *Zorotypus hubbardi* (blind and wingless morph). (*b*) Adult female of *Zorotypus goeleti* (winged morph). (*Modified from D. Grimaldi and M. S. Engel, Evolution of the Insects, Cambridge University Press, 2005; photos © M. S. Engel*)

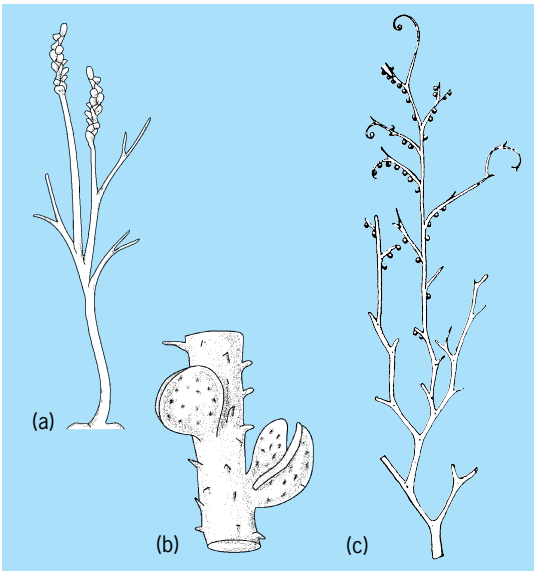
Classification. Zorapterans are notable for having only 32 modern species and six fossil species known, these classified into two genera (*Zorotypus* and *Xenozorotypus*) in a single family. Once believed to be related to termites or booklice, current hypotheses associate the zorapterans with webspinners (Embiodea, Embiidina, Embioptera), as the two groups share a unique pattern of musculature in the hind legs, a reduced anal region in the wings, occurrence of wingless morphs, and a gregarious biology.

Fossil history. Zorapterans are clearly of ancient origin, perhaps diverging from their webspinner relatives during the Late Triassic or Early Jurassic (approximately 200 million years ago). Relatively modern-looking species of Zoraptera are documented in amber from as long ago as the Early Cretaceous, and fossils from this time already exhibit the development of dual morphs within species. *See* INSECTA; ISOPTERA; PSOCOPTERA. Michael S. Engel

Bibliography. J. C. Choe, Courtship feeding and repeated mating in *Zorotypus barberi* (Insecta: Zoraptera) *Anim. Behav.*, 49:1511–1520, 1995; M. S. Engel and D. A. Grimaldi, A winged *Zorotypus* in Miocene amber from the Dominican Republic (Zoraptera: Zorotypidae), with discussion on relationships of and within the order, *Acta Geol. Hisp.*, 35:149–164, 2000; M. S. Engel and D. A. Grimaldi, The first Mesozoic Zoraptera (Insecta), *Amer. Mus. Novit.*, 3362:1–20, 2002; D. Grimaldi and M. S. Engel, *Evolution of the Insects*, 2005; A. B. Gurney, A synopsis of the order Zoraptera, with notes on the biology of *Zorotypus hubbardi Caudell, Proc. Entomol. Soc. Wash.*, 40:57–87, 1938.

Zosterophyllopsida

An extinct class of the division Lycophyta, comprising primitive vascular land plants that evolved from the rhyniophytes and reached a relatively brief peak in the late Early Devonian, when they were an ecologically important and widely distributed element of the exclusively herbaceous terrestrial flora.



Examples of Devonian Zosterophyllopsida. (*a*) Reconstruction of the Early Devonian *Zosterophyllum rhenanum*, here depicted as an aquatic plant (*after R. Kräusel and H. Weyland, Neue Pflanzenfunde im rheinischen Unterdevon, Palaeontographica B, 80:170–190, 1935*). (*b*) Fertile axis of the Middle Devonian *Sawdonia acanthotheca*, showing the unequally valved sporangia and spinelike enations (*after P. G. Gensel, A new species of Sawdonia, Amer. J. Bot., 136:50–62, 1975*). (*c*) Reconstruction of the more derived, Early-Middle Devonian *Gosslingia breconensis* (*after D. Edwards, Fertile Rhyniophytina from the Lower Devonian, Palaeontology, 13:451–461, 1970*).

The Zosterophyllopsida consists of approximately 18 genera, although few of these include species that have been fully reconstructed from their disarticulated organs (see **illus.**). All genera exhibited rhizomatous growth. The rhizomes produced adventitious roots and upright aerial branches, which were typically densely crowded. The group is united by the key lycophyte characters of exarch protoxylem maturation and laterally borne, vascularized, kidney-shaped sporangia with multilayered walls and complete distal dehiscence.

Variation shown by the group is best considered by briefly describing a few of the species most successfully reconstructed from disarticulated organs. *Zosterophyllum llanoveranum* had naked axes up to 6 in. (15 cm) tall, which branched sparsely and somewhat unequally. Its main axes terminated in a reproductive zone that bore many radially arranged sporangia, which terminated short lateral branches. The geographically and stratigraphically more widespread *Sawdonia ornata* was somewhat taller and unequally branched. Globose sporangia were borne on short stalks and were concentrated toward the tips of the more distal lateral branches in approximately linear files. *Gosslingia breconensis* possessed the unornamented axes of *Zosterophyllum* and the size, branching habit, circinate apices, and sporangial arrangement of *Sawdonia*. The sporangia were horizontally oriented, and the stele was elliptical throughout the axes.

The most primitive generally accepted lycophyte is *Zosterophyllum*. *Sawdonia* and *Gosslingia* are more advanced zosterophyllopsids, possessing characters that are absent from the lycopsids. Although briefly successful as dense stands on floodplains and channel margins, these more sophisticated zosterophyllopsids were virtually extinct by the end of the Devonian. *See* LYCOPHYTA; LYCOPSIDA.

Richard M. Bateman; William A. DiMichele

Bibliography. F. M. Hueber, Thoughts on the early lycopsids and zosterophylls, *Ann. Mo. Bot. Gard.*, 79:474–499, 1992; K. J. Niklas and H. P. Banks, A reevaluation of the Zosterophyllophytina, *Amer. J. Bot.*, 77:274–283, 1990; W. N. Stewart and G. W. Rothwell, *Paleobotany and the Evolution of Plants*, 1993.

Zygnematales

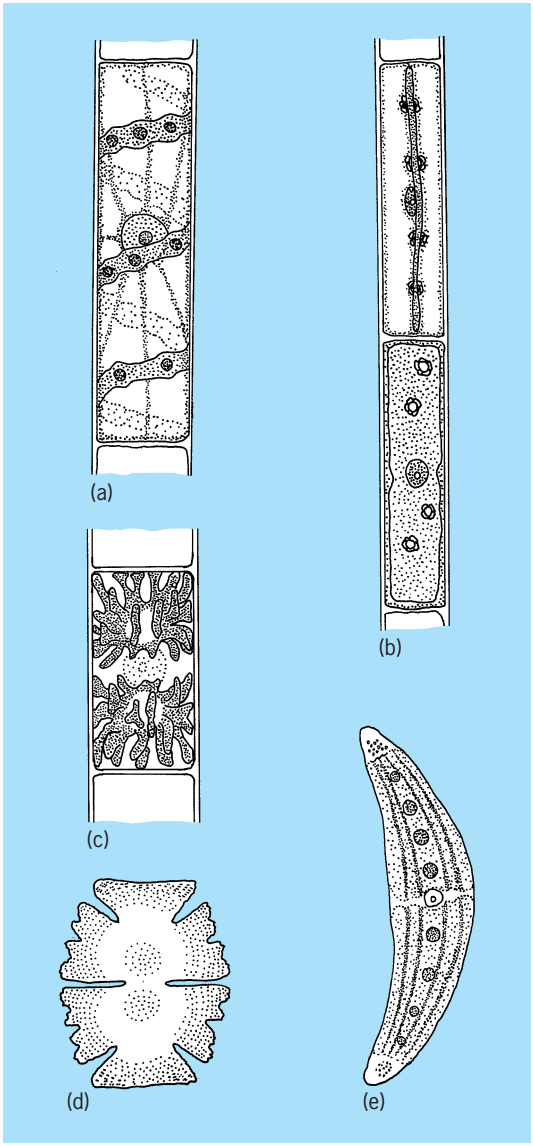
A large order of green algae (Chlorophyceae) that is characterized by the lack of flagellate cells. Sexual reproduction is effected by the fusion of ameboid or passive gametes in conjugation tubes, which give rise to the alternate name Conjugales. Biochemical and ultrastructural features indicate that the Zygnematales are more closely related to charophytes than they are to most other green algae. *See* CHLOROPHYCEAE.

Zygnematales exhibit two somatic types—unicells (desmids) and unbranched uniseriate filaments. The filaments are usually free-floating, but they may be attached by hapteroid outgrowths. The cell wall, which may have pores, is composed of inner fibrillar cellulose layers and an outer slippery pectic layer. All cells are uninucleate. Sexual reproduction is isogamous or anisogamous, with the zygote developing in the conjugation tube or in one of the parents, respectively. The zygote secretes a thick wall that is usually ornamented or sculptured, and undergoes a period of dormancy. Meiosis occurs at some time (variable, depending upon the species) between conjugation and germination.

Members of this order are essentially restricted to fresh-water and subaerial habitats. They are sometimes placed in their own class, with each family elevated to the rank of order. In the present treatment, the order Zygnematales is divided into two suborders, Zygnematineae and Desmidiineae.

Zygnematineae. In this suborder the cells are cylindrical, of uniform diameter, and solitary or in uniseriate filaments. The vegetative cell wall lacks pores or ornamentation. The chloroplast complement of the cell is one of three types: one or more parietal, usually spirally twisted ribbons extending the length of the cell, each with several pyrenoids; an axile plate extending the length of the cell, without pyrenoids or with few to several pyrenoids; and an axile star or disk lying on either side of the central nucleus, each with a pyrenoid. This suborder includes about 18 genera and 650 species.

All filamentous forms are placed in the family Zygnemataceae. In most genera the protoplast of any cell can function directly as a gamete, but in some



Representative genera of Zygnematales. (*a*) *Spirogyra* (family Zygnemataceae). Cell of filament showing spiral chloroplast with pyrenoids and central nucleus. (*b*) *Mougeotia* (Zygnemataceae). Two cells of filament showing axile ribbonlike chloroplast with pyrenoids, edge view in upper cell, face view in lower cell, and central nucleus. (*c*) *Zygnema* (Zygnemataceae). Cell of filament showing pair of stellate chloroplasts, one on either side of central nucleus. (*d*) *Micrasterias* (Desmidiaceae). A placoderm desmid with incised semicells joined by narrow isthmus. (*e*) *Closterium* (Peniaceae). A placoderm desmid without a median constriction.

genera a division prior to conjugation produces a pair of cells, one of which remains vegetative. Conjugation may be between gametes derived from cells of one filament (lateral) or adjacent filaments (scalariform). The entire protoplast of a cell may function as a gamete, or there may be a cytoplasmic residue. Gametes may be isogamous, meeting halfway in a conjugation tube, or anisogamous, with one gamete ameboid and the other passive. Vegetative multiplication by fragmentation is common.

Spirogyra (**illus.** *a*) has spiral chloroplasts and anisogamous conjugation. In *Mougeotia* (illus. *b*) each cell usually has a single axile laminate chloroplast, and conjugation is isogamous. In *Zygnema* (illus. *c*) each cell usually has a pair of stellate chloroplasts, one on either side of the nucleus, and conjugation is isogamous or anisogamous, depending upon the species. *Zygogonium* has a pair of cushion-shaped chloroplasts, at times connected by a bridge. Conjugation is isogamous and involves cytoplasmic residues.

Quiet bodies of fresh water often contain an abundance of the slippery skeins of the common genera of Zygnemataceae, especially *Spirogyra* and *Zygnema*. *Zygogonium* characteristically grows on damp soil.

Members of the Mesotaeniaceae, called saccoderm desmids, are essentially unicellular Zygnemataceae. Following cell division, daughter cells usually separate immediately, but they may remain united temporarily. Spiral chloroplasts are found in *Spirotaenia,* a single axile laminate chloroplast in *Mesotaenium,* and a pair of stellate chloroplasts in *Cylindrocystis*. Conjugation, which usually involves the formation of a tube, is isogamous. Representatives of this family are found in moist subaerial habitats as well as in quiet bodies of fresh water.

Desmidiineae. Members of this suborder, the placoderm desmids, have cells which are variously shaped but always composed of mirror-image semicells demarcated by a median constriction or incision. During division a cell cleaves transversely, and a new semicell is formed by each of the parental semicells. The nucleus lies at the junction of the semicells (where often a sinus constricts the desmid to form an isthmus). There is usually a chloroplast in each semicell, but the form and number of chloroplasts vary in different genera. The cell wall, which is often impregnated with iron salts, has pores and is frequently ornamented. In most genera the cells are solitary, but in some genera they are united in amorphous colonies or unbranched filaments. Locomotion, known in many placoderm desmids, is effected by localized secretion of gelatinous material through pores at one end of the cell. Conjugation takes place within a gelatinous envelope investing the two ameboid protoplasts or, in filamentous forms, within a tube. The zygote secretes a thick wall that is usually ornamented. Placoderm desmids are abundant in fresh-water habitats, especially those that are slightly acid. They often form a film on submerged plants. This suborder comprises about 34 genera and 4000 species.

All desmids with semicells demarcated by a median constriction or incision are included in the family Desmidiaceae. The genera are distinguished primarily on their somatic organization (solitary, colonial, or filamentous), shape and ornamentation of the semicell, and chloroplast features. The most common genera of solitary desmids are *Cosmarium, Euastrum, Micrasterias* (illus. *d*), *Staurastrum, Staurodesmus*, and *Xanthidium*. The intricate lobing of the semicell of *Micrasterias,* coupled with the unusual type of cell division characteristic of the Desmidiaceae, has made this alga a favorite subject for studies in morphogenesis.

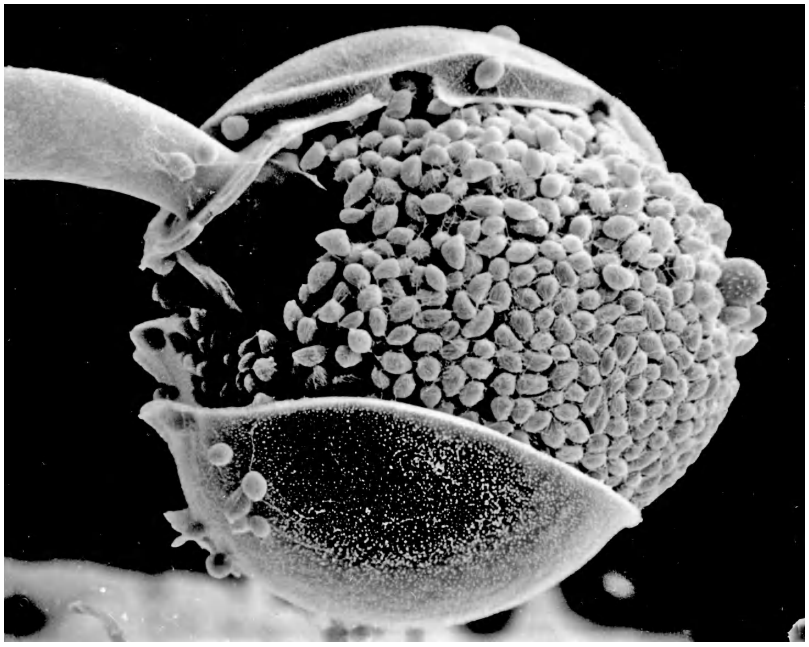
In a small group of placoderm desmids, constituting the family Peniaceae, the cells are cylindrical or lunate, and are composed of semicells not demarcated by a median constriction. Regeneration of a daughter semicell may be followed by elongation of one or both semicells, the zones of elongation being visible in the mature desmid as girdle bands. *Closterium* (illus. *e*) and *Penium* are the most common of the four genera in this family.

Paul C. Silva; Richard L. Moe

Bibliography. H. C. Bold and M. J. Wynne, *Introduction to the Algae: Structure and Reproduction*, 2d ed., 1997; R. E. Lee, *Phycology,* 3d ed., 1999; G. M. Smith, *The Fresh-Water Algae of the United States,* 1950.

Zygomycetes

A class of terrestrial fungi in the phylum Zygomycota, reproducing sexually with a zygospore and asexually by sporangiospore or conidium. The Zygomycetes are one of two classes of the phylum Zygomycota, the other being the arthropod-associated Trichomycetes. Zygomycetes can be distinguished by the production of sporangia (sacs) that undergo internal cleavage to produce multiple sporangiospores (see **illustration**) as opposed to the production of conidiospores as observed in most other groups of fungi (Ascomycota and Basidiomycota). A notable exception is the order Entomophthorales, which reproduces asexually by conidium formation. Zygomycetes are all microscopic with the exception of *Endogone* and relatives that produce a zygospore containing structure called a sporocarp. Zygomycetes are typically studied by first isolating them into pure culture on nutrient media. Sexual reproduction results in the formation of a thick-walled zygosporangium following the conjugation of morphologically relatively similar, undifferentiated gametangia (single-cell structures functioning as gametes). The gametangia may be derived from different (heterothallism) or the same (homothallism) thallus, or genetic individual. *See* EUMYCOTA; FUNGI; ZYGOMYCOTA.



Scanning electron micrograph of dehisced sporangium of *Gilbertella persicaria* releasing sporangiospores. (*Reprinted with permission from K. O'Donnell, Zygomycetes in Culture, 1979;* © *Department of Botany, University of Georgia*)

The Zygomycetes comprise six orders: Dimargaritales, Endogonales, Entomophthorales, Kickxellales, Mucorales, Zoopagales. These orders span a great diversity of ecological habitat. The Dimargaritales and Zoopagales contain many species which parasitize other fungi (mycoparasites). The Entomophthorales are specialized as pathogens primarily of insects. The Kickxellales contains both mycoparasitic species and saprobic species (especially found on soil and animal dung). Endogonales are unique in that some species associate with plant roots as ectomycorrhizae in what is likely to be a mutualistic relationship. The Mucorales is the largest order and are the most commonly encountered species. These species are primarily saprobic (live on decaying organic matter) in dung and soil but are notorious for spoilage of fruits during storage. The Mucorales include some of the fastest-growing fungi, including the famous "hat-thrower" *Pilobolus. Pilobolus* produces sporangia on dung that are forcibly discharged up to 2 m. These sporangia may land on a plant leaf whereupon they are ingested by herbivores and remain viable until being expelled into fresh substrate upon defecation by the animal.

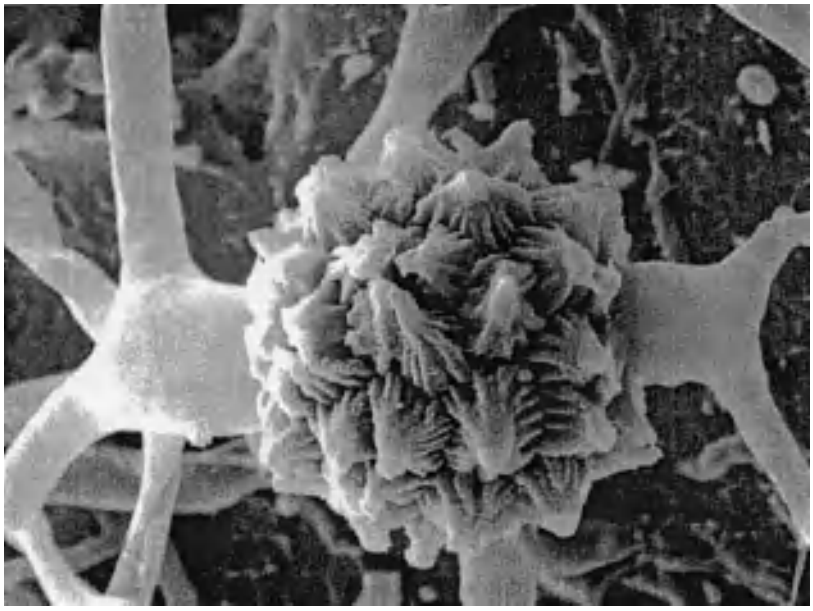
Zygomycetes do not appear to be a natural group. Recent evidence suggests that the Trichomycetes evolved from within the class Zygomycetes, with a strong relationship between the Trichomycetes and Kickxellales.

Timothy James

Bibliography. G. L. Benny, R. A. Humber, and J. B. Morton, Zygomycota: Zygomycetes, in D. J. McLaughlin, E. G. McLaughlin, and P. A. Lemke (eds.), *The Mycota*, vol. 7A: *Systematics and Evolution*, pp. 113–146, Springer Verlag, Berlin, 2001; P. M. Kirk et al. (eds.), *Ainsworth & Bisby's Dictionary of the Fungi*, 9th ed., CAB International, Wallingford, UK, 2001; C. W. Mims, M. M. Blackwell, and C. J. Alexopoulos, *Introductory Mycology*, 4th ed., 1995.

Zygomycota

A phylum of filamentous, microscopic fungi that undergo sexual reproduction by the formation of a zygospore. The Zygomycota are often referred by the common names "pin molds" or "sugar molds" and are ubiquitous fungi though rarely observed by humans because of the diminutive size of their reproductive structures. Zygomycota grow as a network of



Scanning electron micrograph of ornamented zygosporangium of *Micotypha africana*. (*Reprinted with permission from K. O'Donnell, Zygomycetes in Culture, 1979; copyright Department of Botany, University of Georgia*)

interconnected filaments (hyphae) that grow on or inside their substrate and acquire nutrients through absorptive, heterotrophic nutrition. Zygomycota usually have coenocytic (multinucleate) hyphae lacking septa (cross-walls) between cells, but may also be regularly septate. The Zygomycota have a large diversity of ecological roles: as saprophytes on various substrates, especially sugary substrates and dung, as parasites of animals or plants, as commensualistic (both nonpathogenic and nonbeneficial) associates of arthropods, as parasites of other fungi, and as mycorrhizal symbionts with plant roots. *See* MYCORRHIZAE.

As with other fungi, members of the Zygomycota reproduce through spores. Asexual reproduction appears to be the dominant mode of reproduction in most species; sexual reproduction may be rare, or lacking altogether. Asexual reproduction occurs through the production of sporangiospores produced by the internal cleavage of saclike structures termed sporangia. Sporangia are usually borne at the end of a stalk (the sporangiophore) and are dispersed by air or water following the dehiscence (bursting open) of the sporangia wall. Entomophthorales are an exception and reproduce through the production of forcibly discharged conidia. Zygomycetes are haploid throughout nearly all of their life cycle. Sexual reproduction can occur between compatible individuals (heterothallism) or by self-fertilization of a single individual (homothallism). Sex involves the attraction of compatible mates via the hormone trisporic acid. Compatible hyphae grow toward each other in response to a hormonal gradient established by trisporic acid precursors and form specialized cells called gametangia upon contact. The gametangia undergo plasmogamy (fusion) followed by the development of a zygosporangium within which karyogamy (nuclear fusion) takes place, forming the only diploid cell type. Within the zygosporangium, a single zygospore develops. Zygosporangia are typically thick-walled and often contain extensive ornamentation (see **illustration**). Meiosis occurs within the zygosporangium or at the time of germination.

The Zygomycota may be divided into two classes: the Zygomycetes and the Trichomycetes. The Trichomycetes are distinguished from the Zygomycetes by the production of a specialized spore, the trichospore, which is a sporangiospore with an appendage to aid in aquatic dispersal. The Trichomycetes comprise those Zygomycota which have a commensual relationship with arthropods. Until 2000, the Zygomycota also included the arbuscular mycorrhizal fungi which form mutualistic associations with many plant roots. These fungi are now in their own phylum, the Glomeromycota. *See* EUMYCOTA; FUNGI; ZYGOMYCETES.

Bibliography. G. L. Benny, R. A. Humber, and J. B. Morton, Zygomycota: Zygomycetes, in D. J. McLaughlin, E. G. McLaughlin, and P. A. Lemke (eds.), *The Mycota*, vol. 7A, *Systematics and Evolution*, pp. 113–146, Springer, Berlin, 2001; R. W. Lichtwardt, *The Trichomycetes, Fungal Associates of Arthropods*, Springer, New York, 1986; C. W. Mims, M. M. Blackwell, and C. J. Alexopoulos, *Introductory Mycology*, 4th, ed., 1995.

Zygophyllales

An order of flowering plants in the eurosid I group of the rosid eudicots. This order comprises just two small families of distant relationship (as evidenced by DNA sequence studies) to all other rosid eudicots: Krameriaceae and Zygophyllaceae. In previous systems of classification, both families

were placed in the highly unnatural order Polygalales or Sapindales, the families of which share with them only highly zygomorphic flowers. Krameriaceae (25 species) are photosynthetic parasitic plants ("hemiparasites") found from the southwestern United States to South America, whereas Zygophyllaceae (240 species) are free-living and widely distributed in arid to semiarid zones throughout the warmer regions of the world. Some species of the latter family produce especially hard woods such as lignumvitate (*Bulnesia* and *Guaiacum*), and a few others produce medicinal compounds, such as the creosote bush (*Larrea*). *See* FLOWER; LIGNUMVITAE; MAGNOLIOPHYTA; MAGNOLIOPSIDA; POLYGALALES; ROSIDAE; SAPINDALES.

Mark W. Chase